II

BASIC GENETIC MECHANISMS

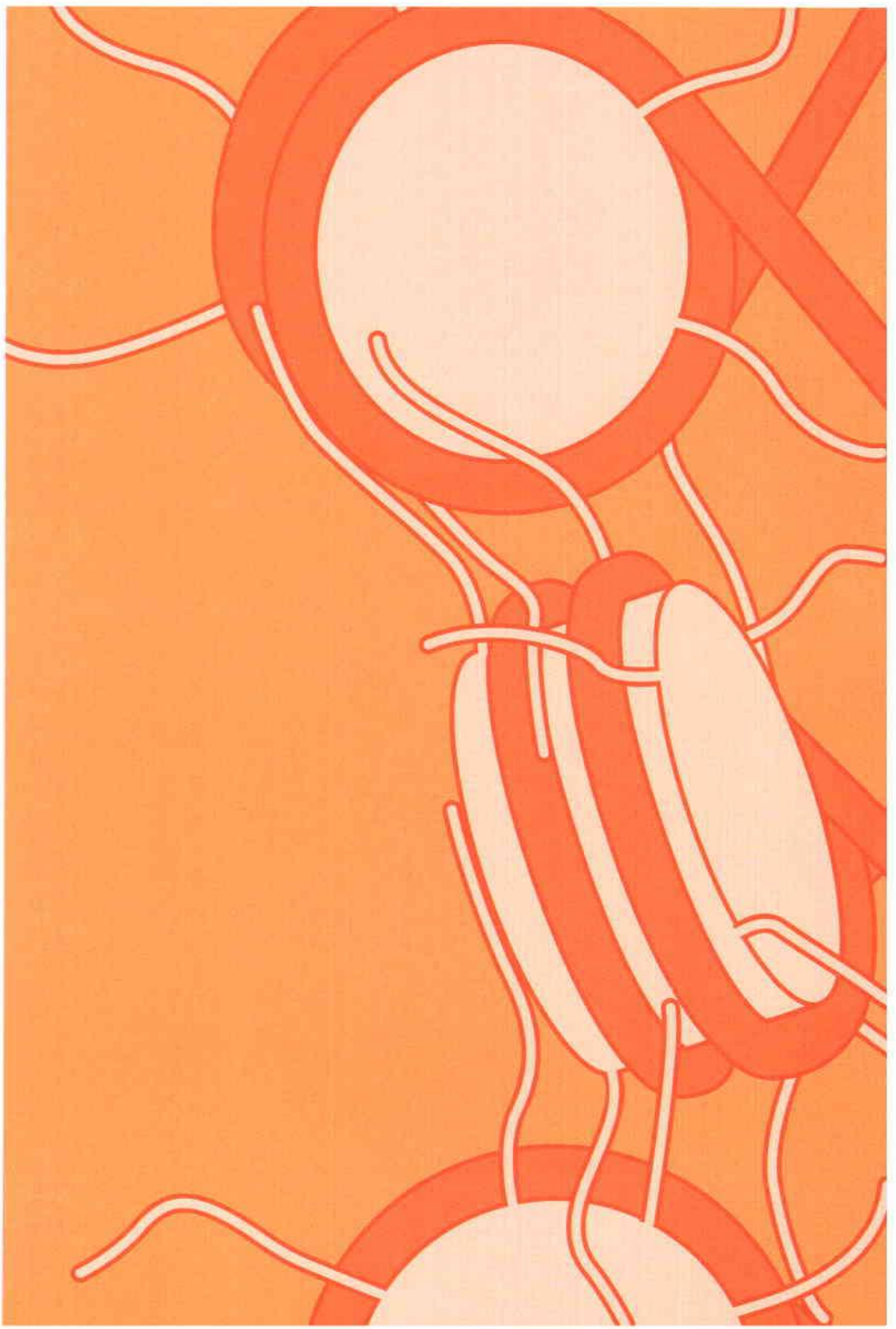
Part II
Chapters

4 DNA, Chromosomes, and Genomes

5 DNA Replication, Repair, and Recombination

6 How Cells Read the Genome: From DNA to Protein

7 Control of Gene Expression



DNA, Chromosomes, and Genomes

4

Life depends on the ability of cells to store, retrieve, and translate the genetic instructions required to make and maintain a living organism. This *hereditary* information is passed on from a cell to its daughter cells at cell division, and from one generation of an organism to the next through the organism's reproductive cells. These instructions are stored within every living cell as its **genes**, the information-containing elements that determine the characteristics of a species as a whole and of the individuals within it.

As soon as genetics emerged as a science at the beginning of the twentieth century, scientists became intrigued by the chemical structure of genes. The information in genes is copied and transmitted from cell to daughter cell millions of times during the life of a multicellular organism, and it survives the process essentially unchanged. What form of molecule could be capable of such accurate and almost unlimited replication and also be able to direct the development of an organism and the daily life of a cell? What kind of instructions does the genetic information contain? How can the enormous amount of information required for the development and maintenance of an organism fit within the tiny space of a cell?

The answers to several of these questions began to emerge in the 1940s. At this time, researchers discovered, from studies in simple fungi, that genetic information consists primarily of instructions for making proteins. Proteins are the macromolecules that perform most cell functions: they serve as building blocks for cell structures and form the enzymes that catalyze the cell's chemical reactions (Chapter 3), they regulate gene expression (Chapter 7), and they enable cells to communicate with each other (Chapter 15) and to move (Chapter 16). The properties and functions of a cell are determined largely by the proteins that it is able to make. With hindsight, it is hard to imagine what other type of instructions the genetic information could have contained.

Painstaking observations of cells and embryos in the late 19th century had led to the recognition that the hereditary information is carried on *chromosomes*, threadlike structures in the nucleus of a eucaryotic cell that become visible by light microscopy as the cell begins to divide (**Figure 4–1**). Later, as biochemical analysis became possible, chromosomes were found to consist of both deoxyribonucleic acid (DNA) and protein. For many decades, the DNA was thought to be merely a structural element. However, the other crucial advance made in the 1940s was the identification of DNA as the likely carrier of genetic information. This breakthrough in our understanding of cells came from studies

In This Chapter

THE STRUCTURE AND FUNCTION OF DNA 197

CHROMOSOMAL DNA AND ITS PACKAGING IN THE CHROMATIN FIBER 202

THE REGULATION OF CHROMATIN STRUCTURE 219

THE GLOBAL STRUCTURE OF CHROMOSOMES 233

HOW GENOMES EVOLVE 245

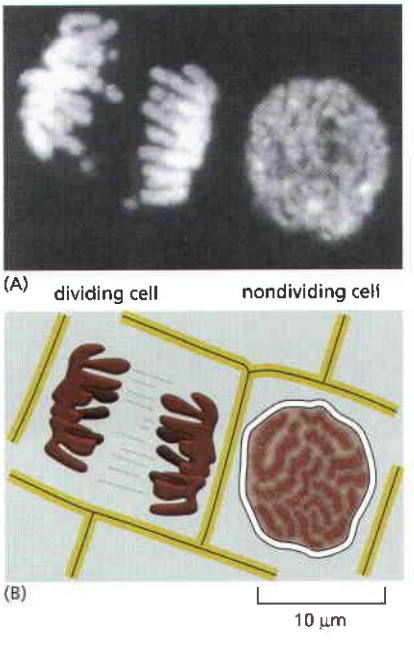


Figure 4–1 Chromosomes in cells. (A) Two adjacent plant cells photographed through a light microscope. The DNA has been stained with a fluorescent dye (DAPI) that binds to it. The DNA is present in chromosomes, which become visible as distinct structures in the light microscope only when they become compact, sausage-shaped structures in preparation for cell division, as shown on the left. The cell on the right, which is not dividing, contains identical chromosomes, but they cannot be clearly distinguished in the light microscope at this phase in the cell's life cycle, because they are in a more extended conformation. (B) Schematic diagram of the outlines of the two cells along with their chromosomes. (A, courtesy of Peter Shaw.)

of inheritance in bacteria (Figure 4–2). But as the 1950s began, both how proteins could be specified by instructions in the DNA and how this information might be copied for transmission from cell to cell seemed completely mysterious. The mystery was suddenly solved in 1953, when the structure of DNA was correctly predicted by James Watson and Francis Crick. As outlined in Chapter 1, the double-helical structure of DNA immediately solved the problem of how the information in this molecule might be copied, or *replicated*. It also provided the first clues as to how a molecule of DNA might use the sequence of its subunits to encode the instructions for making proteins. Today, the fact that DNA is the genetic material is so fundamental to biological thought that it is difficult to appreciate the enormous intellectual gap that was filled.

In this chapter we begin by describing the structure of DNA. We see how, despite its chemical simplicity, the structure and chemical properties of DNA make it ideally suited as the raw material of genes. We then consider how the many proteins in chromosomes arrange and package this DNA. The packing has to be done in an orderly fashion so that the chromosomes can be replicated and apportioned correctly between the two daughter cells at each cell division. It must also allow access to chromosomal DNA for the enzymes that repair it when it is damaged and for the specialized proteins that direct the expression of its many genes. We shall also see how the packaging of DNA differs along the length of each chromosome in eucaryotes, and how it can store a valuable record of the cell's developmental history.

In the past two decades, there has been a revolution in our ability to determine the exact sequence of subunits in DNA molecules. As a result, we now know the order of the 3 billion DNA subunits that provide the information for producing a human adult from a fertilized egg, as well as the DNA sequences of thousands of other organisms. Detailed analyses of these sequences have provided exciting insights into the process of evolution, and it is with this subject that the chapter ends.

This is the first of four chapters that deal with basic genetic mechanisms—the ways in which the cell maintains, replicates, expresses, and occasionally improves the genetic information carried in its DNA. This chapter presents a broad overview of DNA and how it is packaged into chromosomes. In the following chapter (Chapter 5) we discuss the mechanisms by which the cell accurately replicates and repairs DNA; we also describe how DNA sequences can be

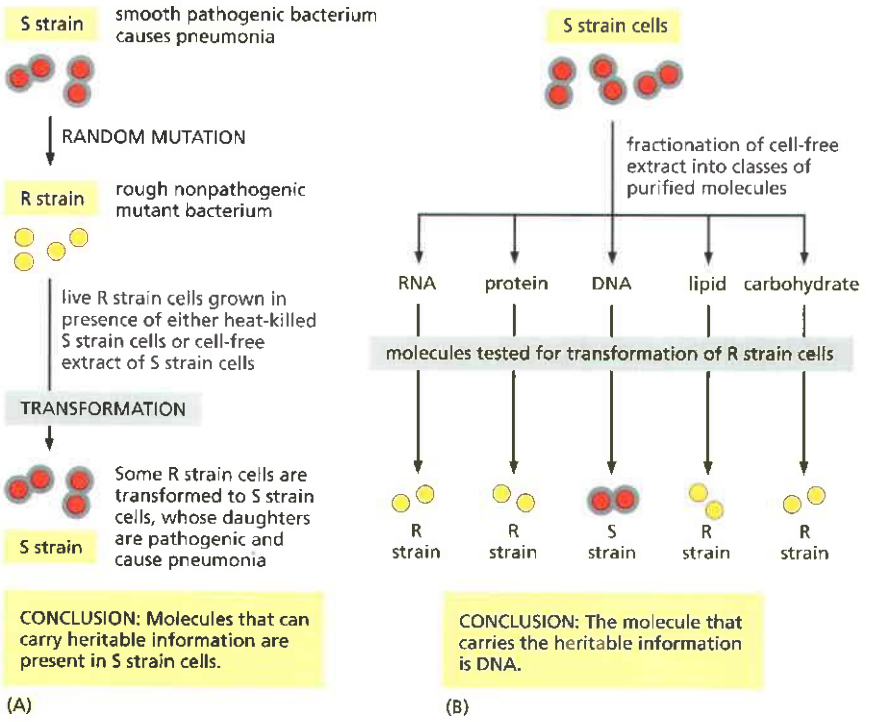


Figure 4–2 The first experimental demonstration that DNA is the genetic material. These experiments, carried out in the 1940s, showed that adding purified DNA to a bacterium changed its properties and that this change was faithfully passed on to subsequent generations. Two closely related strains of the bacterium *Streptococcus pneumoniae* differ from each other in both their appearance under the microscope and their pathogenicity. One strain appears smooth (S) and causes death when injected into mice, and the other appears rough (R) and is nonlethal. (A) An initial experiment shows that a substance present in the S strain can change (or transform) the R strain into the S strain and that this change is inherited by subsequent generations of bacteria. (B) This experiment, in which the R strain has been incubated with various classes of biological molecules purified from the S strain, identifies the substance as DNA.

rearranged through the process of genetic recombination. Gene expression—the process through which the information encoded in DNA is interpreted by the cell to guide the synthesis of proteins—is the main topic of Chapter 6. In Chapter 7, we describe how this gene expression is controlled by the cell to ensure that each of the many thousands of proteins and RNA molecules encrypted in its DNA are manufactured only at the proper time and place in the life of the cell.

THE STRUCTURE AND FUNCTION OF DNA

Biologists in the 1940s had difficulty in conceiving how DNA could be the genetic material because of the apparent simplicity of its chemistry. DNA was known to be a long polymer composed of only four types of subunits, which resemble one another chemically. Early in the 1950s, DNA was examined by x-ray diffraction analysis, a technique for determining the three-dimensional atomic structure of a molecule (discussed in Chapter 8). The early x-ray diffraction results indicated that DNA was composed of two strands of the polymer wound into a helix. The observation that DNA was double-stranded was of crucial significance and provided one of the major clues that led to the Watson–Crick model for DNA structure. But only when this model was proposed in 1953 did DNA's potential for replication and information encoding become apparent. In this section we examine the structure of the DNA molecule and explain in general terms how it is able to store hereditary information.

A DNA Molecule Consists of Two Complementary Chains of Nucleotides

A **deoxyribonucleic acid** (**DNA**) molecule consists of two long polynucleotide chains composed of four types of nucleotide subunits. Each of these chains is known as a *DNA chain*, or a *DNA strand*. *Hydrogen bonds* between the base portions of the nucleotides hold the two chains together (**Figure 4–3**). As we saw in Chapter 2 (Panel 2–6, pp. 116–117), nucleotides are composed of a five-carbon sugar to which are attached one or more phosphate groups and a nitrogen-containing base. In the case of the nucleotides in DNA, the sugar is deoxyribose attached to a single phosphate group (hence the name deoxyribonucleic acid), and the base may be either *adenine (A)*, *cytosine (C)*, *guanine (G)*, or *thymine (T)*. The nucleotides are covalently linked together in a chain through the sugars and phosphates, which thus form a "backbone" of alternating sugar–phosphate–sugar–phosphate. Because only the base differs in each of the four types of subunits, each polynucleotide chain in DNA is analogous to a necklace (the backbone) strung with four types of beads (the four bases A, C, G, and T). These same symbols (A, C, G, and T) are also commonly used to denote the four different nucleotides—that is, the bases with their attached sugar and phosphate groups.

The way in which the nucleotide subunits are linked together gives a DNA strand a chemical polarity. If we think of each sugar as a block with a protruding knob (the 5′ phosphate) on one side and a hole (the 3′ hydroxyl) on the other (see Figure 4–3), each completed chain, formed by interlocking knobs with holes, will have all of its subunits lined up in the same orientation. Moreover, the two ends of the chain will be easily distinguishable, as one has a hole (the 3′ hydroxyl) and the other a knob (the 5′ phosphate) at its terminus. This polarity in a DNA chain is indicated by referring to one end as the *3′ end* and the other as the *5′ end*.

The three-dimensional structure of DNA—the **double helix**—arises from the chemical and structural features of its two polynucleotide chains. Because these two chains are held together by hydrogen bonding between the bases on the different strands, all the bases are on the inside of the double helix, and the sugar-phosphate backbones are on the outside (see Figure 4–3). In each case, a bulkier two-ring base (a purine; see Panel 2–6, pp. 116–117) is paired with a single-ring base (a pyrimidine); A always pairs with T, and G with C (**Figure**

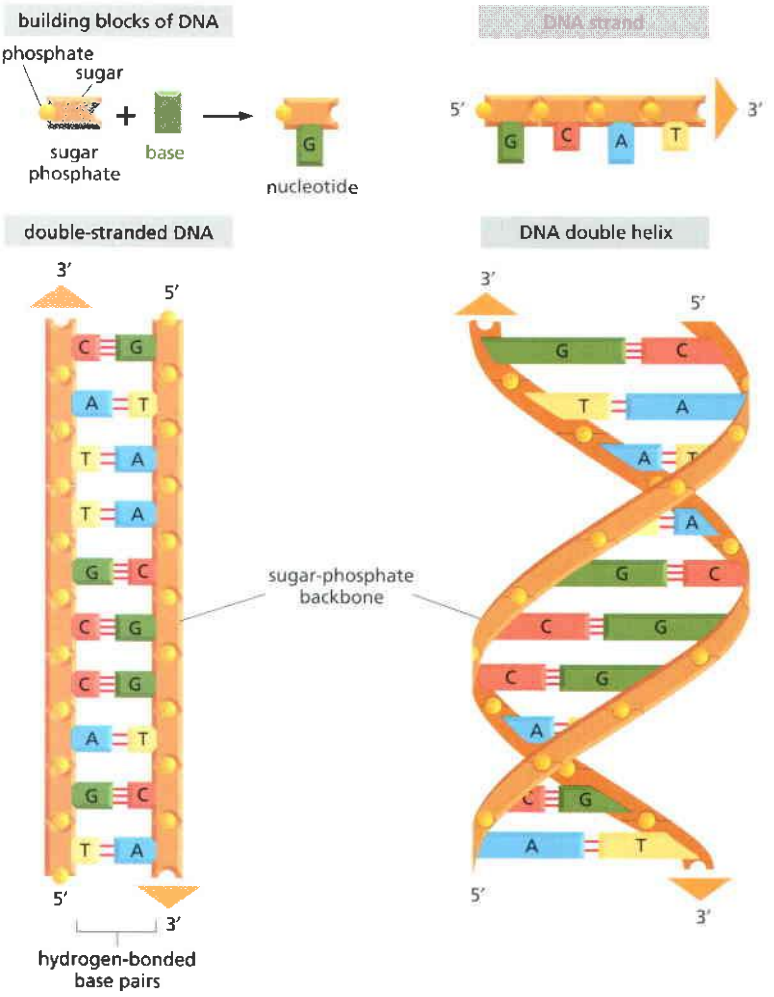


Figure 4–3 DNA and its building blocks. <CAGA> DNA is made of four types of nucleotides, which are linked covalently into a polynucleotide chain (a DNA strand) with a sugar-phosphate backbone from which the bases (A, C, G, and T) extend. A DNA molecule is composed of two DNA strands held together by hydrogen bonds between the paired bases. The *arrowheads* at the ends of the DNA strands indicate the polarities of the two strands, which run antiparallel to each other in the DNA molecule. In the diagram at the bottom left of the figure, the DNA molecule is shown straightened out; in reality, it is twisted into a double helix, as shown on the right. For details, see Figure 4–5.

4–4). This *complementary base-pairing* enables the **base pairs** to be packed in the energetically most favorable arrangement in the interior of the double helix. In this arrangement, each base pair is of similar width, thus holding the sugar-phosphate backbones an equal distance apart along the DNA molecule. To maximize the efficiency of base-pair packing, the two sugar-phosphate backbones

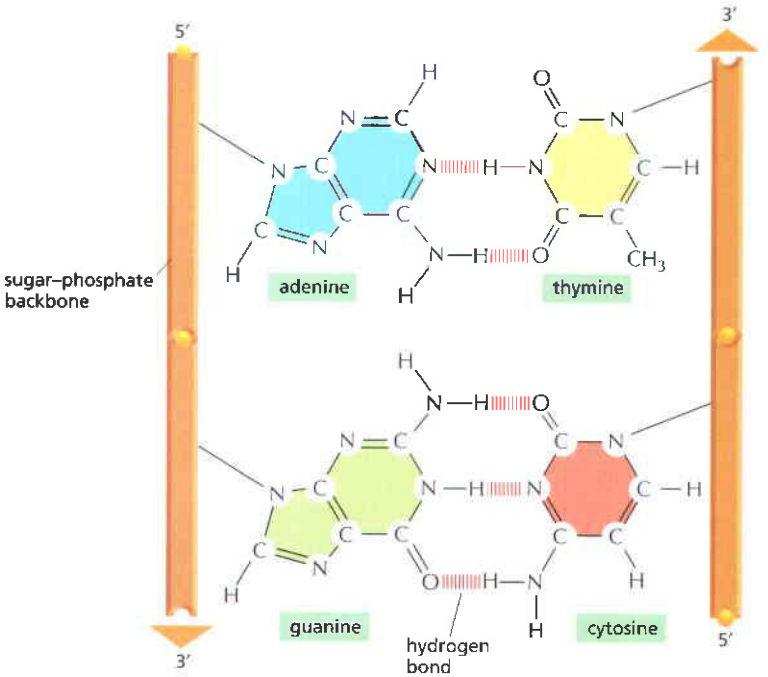


Figure 4–4 Complementary base pairs in the DNA double helix. The shapes and chemical structure of the bases allow hydrogen bonds to form efficiently only between A and T and between G and C, where atoms that are able to form hydrogen bonds (see Panel 2–3, pp. 110–111) can be brought close together without distorting the double helix. As indicated, two hydrogen bonds form between A and T, while three form between G and C. The bases can pair in this way only if the two polynucleotide chains that contain them are antiparallel to each other.

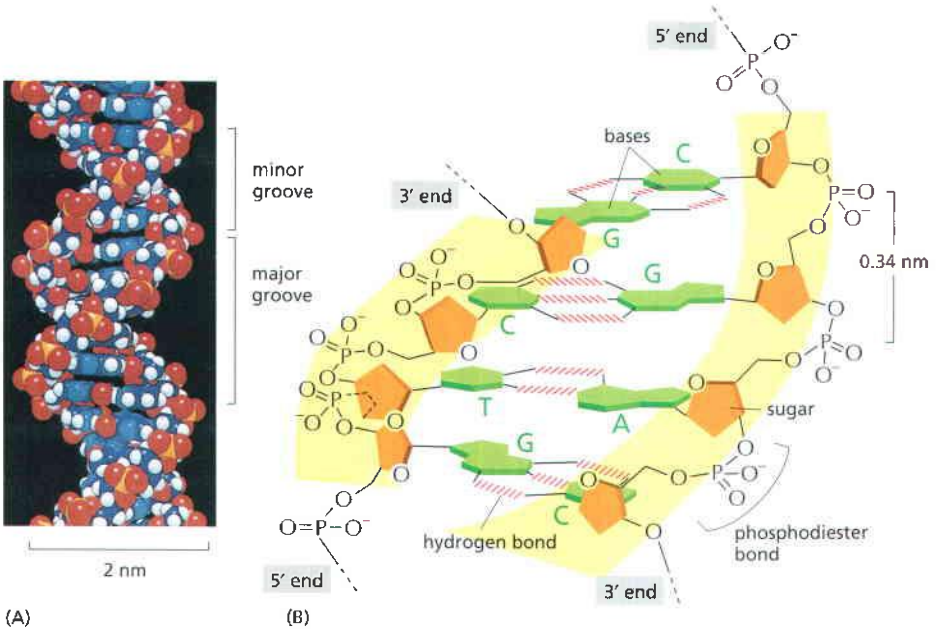


Figure 4–5 The DNA double helix. (A) A space-filling model of 1.5 turns of the DNA double helix. Each turn of DNA is made up of 10.4 nucleotide pairs, and the center-to-center distance between adjacent nucleotide pairs is 3.4 nm. The coiling of the two strands around each other creates two grooves in the double helix: the wider groove is called the major groove, and the smaller the minor groove. (B) A short section of the double helix viewed from its side, showing four base pairs. The nucleotides are linked together covalently by phosphodiester bonds that join the 3′-hydroxyl (–OH) group of one sugar to the 5′-hydroxyl group of the next sugar. Thus, each polynucleotide strand has a chemical polarity; that is, its two ends are chemically different. The 5′ end of the DNA polymer is by convention often illustrated carrying a phosphate group, while the 3′-end is shown with a hydroxyl.

wind around each other to form a double helix, with one complete turn every ten base pairs (Figure 4–5).

The members of each base pair can fit together within the double helix only if the two strands of the helix are **antiparallel**—that is, only if the polarity of one strand is oriented opposite to that of the other strand (see Figures 4–3 and 4–4). A consequence of these base-pairing requirements is that each strand of a DNA molecule contains a sequence of nucleotides that is exactly **complementary** to the nucleotide sequence of its partner strand.

The Structure of DNA Provides a Mechanism for Heredity

Genes carry biological information that must be copied accurately for transmission to the next generation each time a cell divides to form two daughter cells. Two central biological questions arise from these requirements: how can the information for specifying an organism be carried in chemical form, and how is it accurately copied? The discovery of the structure of the DNA double helix was a landmark in twentieth-century biology because it immediately suggested answers to both questions, thereby providing a molecular explanation for the problem of heredity. We discuss these answers briefly in this section, and we shall examine them in much more detail in subsequent chapters.

DNA encodes information through the order, or sequence, of the nucleotides along each strand. Each base—A, C, T, or G—can be considered as a letter in a four-letter alphabet that spells out biological messages in the chemical structure of the DNA. As we saw in Chapter 1, organisms differ from one another because their respective DNA molecules have different nucleotide sequences and, consequently, carry different biological messages. But how is the nucleotide alphabet used to make messages, and what do they spell out?

As discussed above, it was known well before the structure of DNA was determined that genes contain the instructions for producing proteins. The DNA messages must therefore somehow encode proteins (Figure 4–6). This relationship immediately makes the problem easier to understand. As discussed in Chapter 3, the properties of a protein, which are responsible for its biological function, are determined by its three-dimensional structure. This structure is determined in turn by the linear sequence of the amino acids of which it is composed. The linear sequence of nucleotides in a gene must therefore somehow spell out the linear sequence of amino acids in a protein. The exact correspondence between the four-letter nucleotide alphabet of DNA and the twenty-letter amino acid alphabet of proteins—the genetic code—is not obvious from the DNA structure, and it took over a decade after the discovery of the double helix

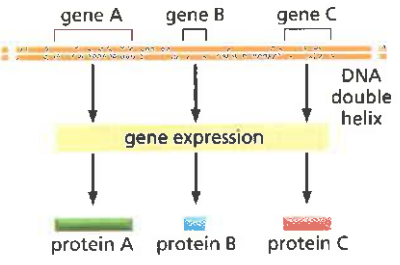


Figure 4–6 The relationship between genetic information carried in DNA and proteins (discussed in Chapter 1).

before it was worked out. In Chapter 6 we will describe this code in detail in the course of elaborating the process, known as *gene expression*, through which a cell converts the nucleotide sequence of a gene first into the nucleotide sequence of an RNA molecule, and then into the amino acid sequence of a protein.

The complete set of information in an organism's DNA is called its **genome**, and it carries the information for all the proteins and RNA molecules that the organism will ever synthesize. (The term genome is also used to describe the DNA that carries this information.) The amount of information contained in genomes is staggering: for example, a typical human diploid cell contains 2 meters of DNA double helix. Written out in the four-letter nucleotide alphabet, the nucleotide sequence of a very small human gene occupies a quarter of a page of text (**Figure 4–7**), while the complete sequence of nucleotides in the human genome would fill more than a thousand books the size of this one. In addition to other critical information, it carries the instructions for roughly 24,000 distinct proteins.

At each cell division, the cell must copy its genome to pass it to both daughter cells. The discovery of the structure of DNA also revealed the principle that makes this copying possible: because each strand of DNA contains a sequence of nucleotides that is exactly complementary to the nucleotide sequence of its partner strand, each strand can act as a **template**, or mold, for the synthesis of a new complementary strand. In other words, if we designate the two DNA strands as S and S′, strand S can serve as a template for making a new strand S′, while strand S′ can serve as a template for making a new strand S (**Figure 4–8**). Thus, the genetic information in DNA can be accurately copied by the beautifully simple process in which strand S separates from strand S′, and each separated strand then serves as a template for the production of a new complementary partner strand that is identical to its former partner.

The ability of each strand of a DNA molecule to act as a template for producing a complementary strand enables a cell to copy, or *replicate*, its genome before passing it on to its descendants. In the next chapter we shall describe the elegant machinery the cell uses to perform this enormous task.

In Eucaryotes, DNA Is Enclosed in a Cell Nucleus

As described in Chapter 1, nearly all the DNA in a eucaryotic cell is sequestered in a nucleus, which in many cells occupies about 10% of the total cell volume. This compartment is delimited by a *nuclear envelope* formed by two concentric lipid bilayer membranes (**Figure 4–9**). These membranes are punctured at intervals by large nuclear pores, which transport molecules between the nucleus and the cytosol. The nuclear envelope is directly connected to the extensive membranes of the endoplasmic reticulum, which extend out from it into the cytoplasm. And it is mechanically supported by a network of intermediate filaments called the *nuclear lamina*, which forms a thin sheetlike meshwork just beneath the inner nuclear membrane (see Figure 4–9B).

The nuclear envelope allows the many proteins that act on DNA to be concentrated where they are needed in the cell, and, as we see in subsequent

CCCTGTGGAGCCACACCCTAGGGTTGGCCA
ATCTACTCCCAGGAGCAGGGAGGGCAGGAG
CCAGGGCTGGGCATAAAAGTCAGGGCAGAG
CCATCTATTGCTTACATTTGCTTCTGACAC
AACTGTGTTCACTAGCAACTCAAACAGACA
CCATGGTGCACCTGACTCCTGAGGAGAAGT
CTGCCGTTACTGCCCTGTGGGGCAAGGTGA
ACGTGGATGAAGTTGGTGGTGAGGCCCTGG
GCAGGTTGGTATCAAGGTTACAAGACAGGT
TTAAGGAGACCAATAGAAACTGGGCATGTG
GAGACAGAGAAGACTCTTGGGTTTCTGATA
GGCACTGACTCTCTCTGCCTATTGGTCTAT
TTTCCCACCCTTAGGCTGCTGGTGGTCTAC
CCTTGGACCCAGAGGTTCTTTGAGTCCTTT
GGGGATCTGTCCACTCCTGATGCTGTTATG
GGCAACCCTAAGGTGAAGGCTCATGGCAAG
AAAGTGCTCGGTGCCTTTAGTGATGGCCTG
GCTCACCTGGACAACCTCAAGGGCACCTTT
GCCACACTGAGTGAGCTGCACTGTGACAAG
CTGCACGTGGATCCTGAGAACTTCAGGGTG
AGTCTATGGGACCCTTGATGTTTTCTTTCC
CCTTCTTTTCTATGGTTAAGTTCATGTCAT
AGGAAGGGGAGAAGTAACAGGGTACAGTTT
AGAATGGGAAACAGACGAATGATTGCATCA
GTGTGGAAGTCTCAGGATCGTTTTAGTTTC
TTTTATTTGCTGTTCATAACAATTGTTTTC
TTTTGTTTAATTCTTGCTTTCTTTTTTTTT
CTTCTCCGCAATTTTTACTATTATACTTAA
TGCCTTAACATTGTGTATAACAAAAGGAAA
TATCTCTGAGATACATTAAGTAACTTAAAA
AAAAACTTTACACAGTCTGCCTAGTACATT
ACTATTTGGAATATATGTGTGCTTATTTGC
ATATTCATAATCTCCCTACTTTATTTTCTT
TTATTTTTAATTGATACATAATCATTATAC
ATATTTATGGGTTAAAGTGTAATGTTTTAA
TATGTGTACACATATTGACCAAATCAGGGT
AATTTTGCATTTGTAATTTTAAAAAATGCT
TTCTTCTTTTAATATACTTTTTTGTTTATC
TTATTTCTAATACTTTCCCTAATCTCTTTC
TTTCAGGGCAATAATGATACAATGTATCAT
GCCTCTTTGCACCATTCTAAAGAATAACAG
TGATAATTTCTGGGTTAAGGCAATAGCAAT
ATTTCTGCATATAAATATTTCTGCATATAA
ATTGTAACTGATGTAAGAGGTTTCATATTG
CTAATAGCAGCTACAATCCAGCTACCATTC
TGCTTTTATTTTATGGTTGGGATAAGGCTG
GATTATTCTGAGTCCAAGCTAGGCCCTTTT
GCTAATCATGTTCATACCTCTTATCTTCCT
CCCACAGCTCCTGGGCAACGTGCTGGTCTG
TGTGCTGGCCCATCACTTTGGCAAAGAATT
CACCCCACCAGTGCAGGCTGCCTATCAGAA
AGTGGTGGCTGGTGTGGCTAATGCCCTGGC
CCACAAGTATCACTAAGCTCGCTTTCTTGC
TGTCCAATTTCTATTAAAGGTTCCTTTGTT
CCCTAAGTCCAACTACTAAACTGGGGGATA
TTATGAAGGGCCTTGAGCATCTGGATTCTG
CCTAATAAAAAACATTTATTTTCATTGCAA
TGATGTATTTAAATTATTTCTGAATATTTT
ACTAAAAAGGGAATGTGGGAGGTCAGTGCA
TTTAAAACATAAAGAAATGATGAGCTGTTC
AAACCTTGGGAAAATACACTATATCTTAAA
CTCCATGAAAGAAGGTGAGGCTGCAACCAG
CTAATGCACATTGGCAACAGCCCCTGATGC
CTATGCCTTATTCATCCCTCAGAAAAGGAT
TCTTGTAGAGGCTTGATTTGCAGGTTAAAG
TTTTGCTATGCTGTATTTTACATTACTTAT
TGTTTTAGCTGTCCTCATGAATGTCTTTTC

Figure 4–7 The nucleotide sequence of the human β-globin gene. By convention, a nucleotide sequence is written from its 5′ end to its 3′ end, and it should be read from left to right in successive lines down the page as though it were normal English text. This gene carries the information for the amino acid sequence of one of the two types of subunits of the hemoglobin molecule, the protein that carries oxygen in the blood. A different gene, the α-globin gene, carries the information for the other type of hemoglobin subunit (a hemoglobin molecule has four subunits, two of each type). Only one of the two strands of the DNA double helix containing the β-globin gene is shown; the other strand has the exact complementary sequence. The DNA sequences highlighted in *yellow* show the three regions of the gene that specify the amino acid sequence for the β-globin protein. We shall see in Chapter 6 how the cell splices these three sequences together at the level of messenger RNA in order to synthesize a full-length β-globin protein.

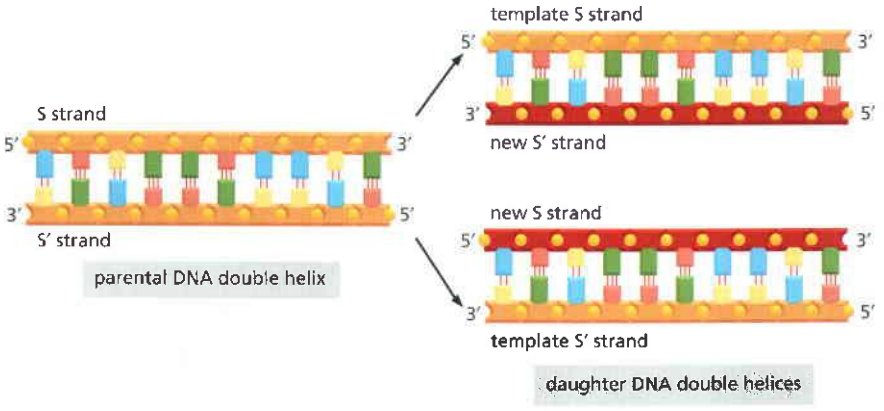


Figure 4–8 DNA as a template for its own duplication. As the nucleotide A successfully pairs only with T, and G with C, each strand of DNA can act as a template to specify the sequence of nucleotides in its complementary strand. In this way, double-helical DNA can be copied precisely, with each parental DNA helix producing two identical daughter DNA helices.

chapters, it also keeps nuclear and cytosolic enzymes separate, a feature that is crucial for the proper functioning of eucaryotic cells. Compartmentalization, of which the nucleus is an example, is an important principle of biology; it serves to establish an environment in which biochemical reactions are facilitated by the high concentration of both substrates and the enzymes that act on them. Compartmentalization also prevents enzymes needed in one part of the cell from interfering with the orderly biochemical pathways in another.

Summary

Genetic information is carried in the linear sequence of nucleotides in DNA. Each molecule of DNA is a double helix formed from two complementary strands of nucleotides held together by hydrogen bonds between G-C and A-T base pairs. Duplication of the genetic information occurs by the use of one DNA strand as a template for the formation of a complementary strand. The genetic information stored in an organism's DNA contains the instructions for all the proteins the organism will ever synthesize and is said to comprise its genome. In eucaryotes, DNA is contained in the cell nucleus, a large membrane-bound compartment.

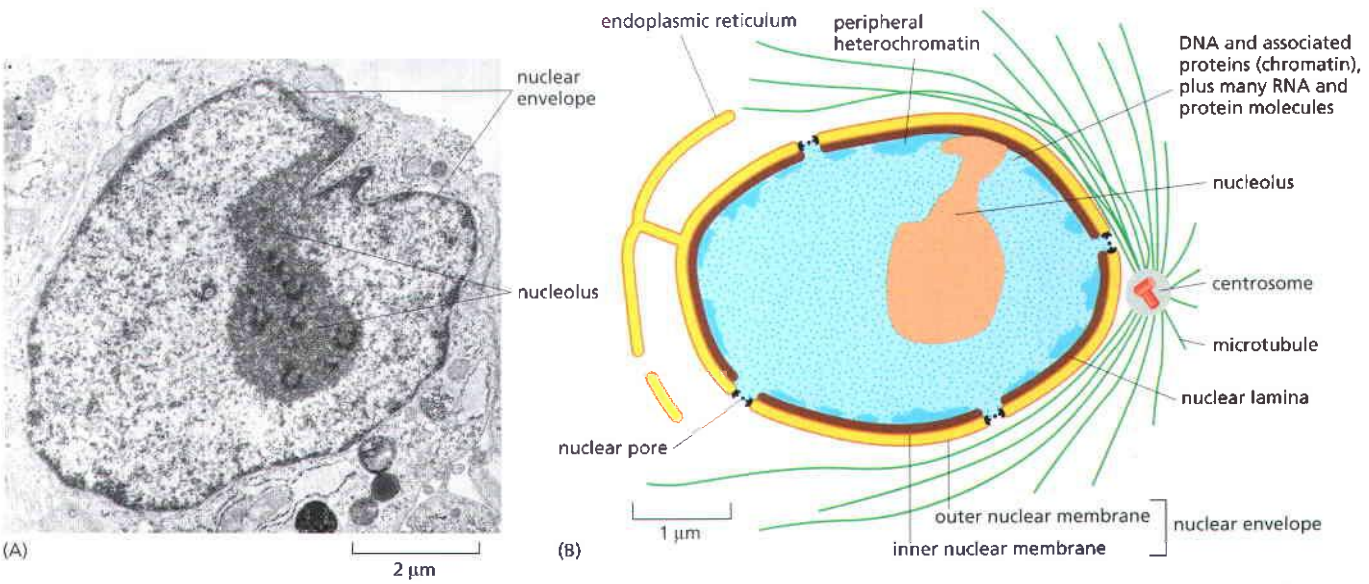


Figure 4–9 A cross-sectional view of a typical cell nucleus. (A) Electron micrograph of a thin section through the nucleus of a human fibroblast. (B) Schematic drawing, showing that the nuclear envelope consists of two membranes, the outer one being continuous with the endoplasmic reticulum membrane (see also Figure 12–8). The space inside the endoplasmic reticulum (the ER lumen) is colored *yellow;* it is continuous with the space between the two nuclear membranes. The lipid bilayers of the inner and outer nuclear membranes are connected at each nuclear pore. A sheet-like network of intermediate filaments *(brown)* inside the nucleus provides mechanical support for the nuclear envelope, forming a special supporting structure called the nuclear lamina (for details, see Chapter 12). The heterochromatin near the lamina contains specially condensed regions of DNA that will be discussed later.

CHROMOSOMAL DNA AND ITS PACKAGING IN THE CHROMATIN FIBER

The most important function of DNA is to carry genes, the information that specifies all the proteins and RNA molecules that make up an organism—including information about when, in what types of cells, and in what quantity each protein is to be made. The genomes of eucaryotes are divided up into chromosomes, and in this section we see how genes are typically arranged on each chromosome. In addition, we describe the specialized DNA sequences that are required for a chromosome to be accurately duplicated and passed on from one generation to the next.

We also confront the serious challenge of DNA packaging. If the double helices comprising all 46 chromosomes in a human cell could be laid end-to-end, they would reach approximately 2 meters; yet the nucleus, which contains the DNA, is only about 6 μm in diameter. This is geometrically equivalent to packing 40 km (24 miles) of extremely fine thread into a tennis ball! The complex task of packaging DNA is accomplished by specialized proteins that bind to and fold the DNA, generating a series of coils and loops that provide increasingly higher levels of organization, preventing the DNA from becoming an unmanageable tangle. Amazingly, although the DNA is very tightly folded, it is compacted in a way that keeps it available to the many enzymes in the cell that replicate it, repair it, and use its genes to produce RNA molecules and proteins.

Eucaryotic DNA Is Packaged into a Set of Chromosomes

In eucaryotes, the DNA in the nucleus is divided between a set of different **chromosomes**. For example, the human genome—approximately 3.2×10^9 nucleotides—is distributed over 24 different chromosomes. Each chromosome consists of a single, enormously long linear DNA molecule associated with proteins that fold and pack the fine DNA thread into a more compact structure. The complex of DNA and protein is called *chromatin* (from the Greek *chroma*, "color," because of its staining properties). In addition to the proteins involved in packaging the DNA, chromosomes are also associated with many proteins and RNA molecules required for the processes of gene expression, DNA replication, and DNA repair.

Bacteria carry their genes on a single DNA molecule, which is often circular (see Figure 1–29). This DNA is associated with proteins that package and condense the DNA, but they are different from the proteins that perform these functions in eucaryotes. Although often called the bacterial "chromosome," it does not have the same structure as eucaryotic chromosomes, and less is known about how the bacterial DNA is packaged. Therefore, our discussion of chromosome structure will focus almost entirely on eucaryotic chromosomes.

With the exception of the germ cells (discussed in Chapter 21) and a few highly specialized cell types that cannot multiply and lack DNA altogether (for example, red blood cells), each human cell contains two copies of each chromosome, one inherited from the mother and one from the father. The maternal and paternal chromosomes of a pair are called **homologous chromosomes** (**homologs**). The only nonhomologous chromosome pairs are the sex chromosomes in males, where a *Y chromosome* is inherited from the father and an *X chromosome* from the mother. Thus, each human cell contains a total of 46 chromosomes—22 pairs common to both males and females, plus two so-called sex chromosomes (X and Y in males, two Xs in females). *DNA hybridization* is a technique in which a labeled nucleic acid strand serves as a "probe" that localizes a complementary strand, as will be described in detail in Chapter 8. This technique can be used to distinguish these human chromosomes by "painting" each one a different color (**Figure 4–10**). Chromosome painting is typically done at the stage in the cell cycle called mitosis, when chromosomes are especially compacted and easy to visualize (see below).

Another more traditional way to distinguish one chromosome from another

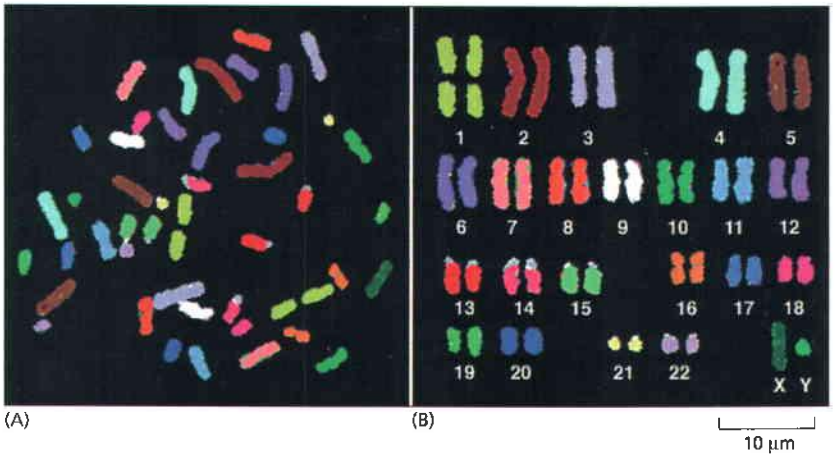


Figure 4–10 The complete set of human chromosomes. These chromosomes, from a male, were isolated from a cell undergoing nuclear division (mitosis) and are therefore highly compacted. Each chromosome has been "painted" a different color to permit its unambiguous identification under the light microscope. Chromosome painting is performed by exposing the chromosomes to a collection of human DNA molecules that have been coupled to a combination of fluorescent dyes. For example, DNA molecules derived from chromosome 1 are labeled with one specific dye combination, those from chromosome 2 with another, and so on. Because the labeled DNA can form base pairs, or hybridize, only to the chromosome from which it was derived (discussed in Chapter 8), each chromosome is differently labeled. For such experiments, the chromosomes are subjected to treatments that separate the double-helical DNA into individual strands, designed to permit base-pairing with the single-stranded labeled DNA while keeping the chromosome structure relatively intact. (A) The chromosomes visualized as they originally spilled from the lysed cell. (B) The same chromosomes artificially lined up in their numerical order. This arrangement of the full chromosome set is called a karyotype. (From E. Schröck et al., *Science* 273:494–497, 1996. With permission from AAAS.)

along each mitotic chromosome (**Figure 4–11**). The structural bases for these banding patterns are not well understood. Nevertheless, the pattern of bands on each type of chromosome is unique, and it is these patterns that initially allowed each human chromosome to be identified and numbered.

The display of the 46 human chromosomes at mitosis is called the human **karyotype**. If parts of chromosomes are lost or are switched between chromosomes, these changes can be detected by changes in the banding patterns or by changes in the pattern of chromosome painting (**Figure 4–12**). Cytogeneticists use these alterations to detect chromosome abnormalities that are associated with inherited defects, as well as to characterize cancers that are associated with specific chromosome rearrangements in somatic cells (discussed in Chapter 20).

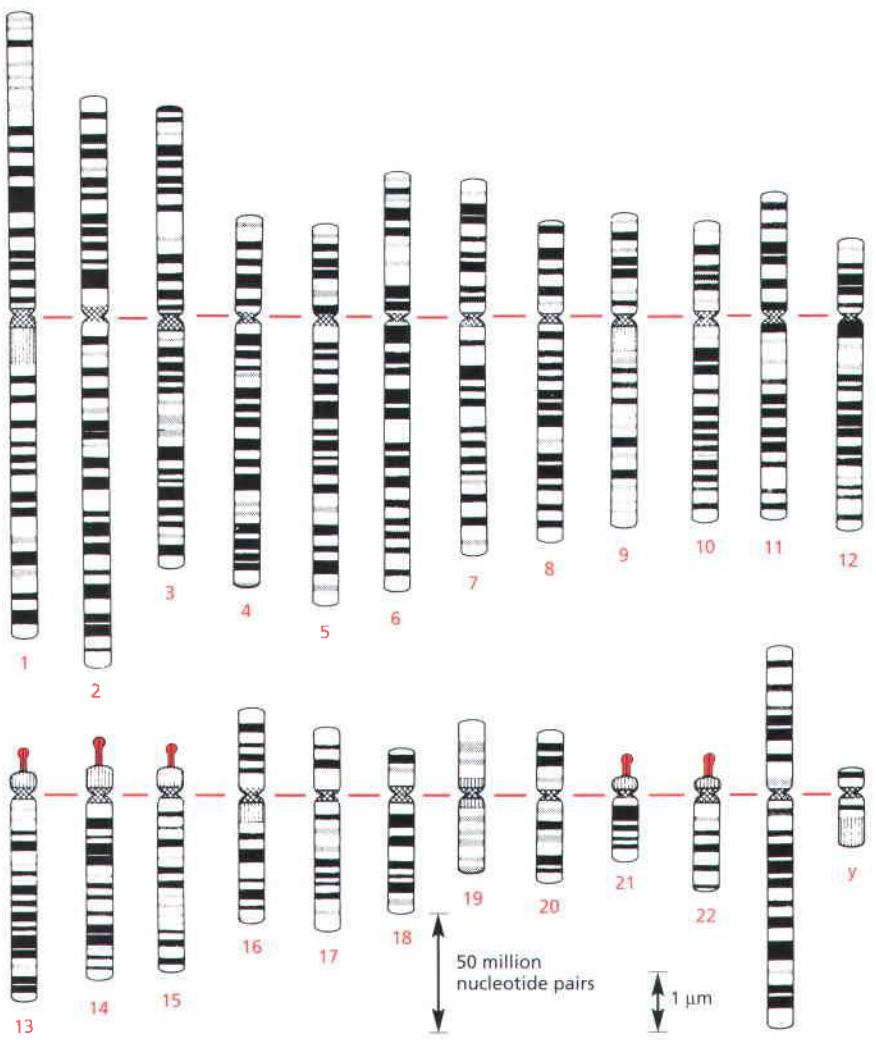


Figure 4–11 The banding patterns of human chromosomes. Chromosomes 1–22 are numbered in approximate order of size. A typical human somatic (non-germ-line) cell contains two of each of these chromosomes, plus two sex chromosomes—two X chromosomes in a female, one X and one Y chromosome in a male. The chromosomes used to make these maps were stained at an early stage in mitosis, when the chromosomes are incompletely compacted. The *horizontal red line* represents the position of the centromere (see Figure 4–21), which appears as a constriction on mitotic chromosomes. The red knobs on chromosomes 13, 14, 15, 21, and 22 indicate the positions of genes that code for the large ribosomal RNAs (discussed in Chapter 6). These patterns are obtained by staining chromosomes with Giemsa stain, and they can be observed under the light microscope. (For micrographs, see Figure 21–18; adapted from U. Franke, *Cytogenet. Cell Genet.* 31:24–32, 1981. With

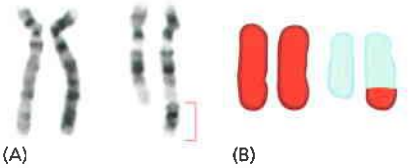


Figure 4–12 An aberrant human chromosome. (A) Two pairs of chromosomes, stained with Giemsa (see Figure 4–11), from a patient with ataxia, a disease characterized by progressive deterioration of motor skills. The patient has a normal pair of chromosome 4s *(left-hand pair)*, but one normal chromosome 12 and one aberrant chromosome 12, as seen by its greater length *(right-hand pair)*. The additional material contained on the aberrant chromosome 12 (*red bracket*) was deduced, from its pattern of bands, as a copy of part of chromosome 4 that had become attached to chromosome 12 through an abnormal recombination event, called a chromosomal translocation. (B) Drawing of the same two chromosome pairs, "painted" *red* for chromosome 4 DNA and *blue* for chromosome 12 DNA. The two techniques give rise to the same conclusion regarding the nature of the aberrant chromosome 12, but chromosome painting provides better resolution, allowing the clear identification of even short pieces of chromosomes that have become translocated. However, Giemsa staining is easier to perform. (Adapted from E. Schröck et al., *Science* 273:494–497, 1996. With permission from AAAS.)

Chromosomes Contain Long Strings of Genes

Chromosomes carry genes—the functional units of heredity. A gene is usually defined as a segment of DNA that contains the instructions for making a particular protein (or a set of closely related proteins). Although this definition holds for the majority of genes, several percent of genes produce an RNA molecule, instead of a protein, as their final product. Like proteins, these RNA molecules perform a diverse set of structural and catalytic functions in the cell, and we discuss them in detail in subsequent chapters.

As might be expected, some correlation exists between the complexity of an organism and the number of genes in its genome (see Table 1–1, p. 18). For example, some simple bacteria have only 500 genes, compared to about 25,000 for humans. Bacteria and some single-celled eucaryotes, such as yeast, have especially concise genomes; the complete nucleotide sequence of their genomes reveals that the DNA molecules that make up their chromosomes are little more than strings of closely packed genes (**Figure 4–13**). However, chromosomes from many eucaryotes (including humans) contain, in addition to genes, a large excess of interspersed DNA that does not seem to carry critical information. Sometimes called "junk DNA" to signify that its usefulness to the cell has not been demonstrated, the particular nucleotide sequence of most of this DNA may not be important. However, some of this DNA is crucial for the proper expression of certain genes, as we discuss elsewhere.

Because of differences in the amount of DNA interspersed between genes, genome sizes can vary widely (see Figure 1–37). For example, the human genome is 200 times larger than that of the yeast *S. cerevisiae*, but 30 times smaller than that of some plants and amphibians and 200 times smaller than that of a species of amoeba. Moreover, because of differences in the amount of excess DNA, the genomes of similar organisms (bony fish, for example) can vary several hundredfold in their DNA content, even though they contain roughly the same number of genes. Whatever the excess DNA may do, it seems clear that it is not a great handicap for a eucaryotic cell to carry a large amount of it.

How the genome is divided into chromosomes also differs from one eucaryotic species to the next. For example, compared with 46 for humans, somatic cells from a species of small deer contain only 6 chromosomes, while those from a species of carp contain over 100. Even closely related species with similar genome sizes can have very different numbers and sizes of chromosomes (**Figure 4–14**). Thus, there is no simple relationship between chromosome number,

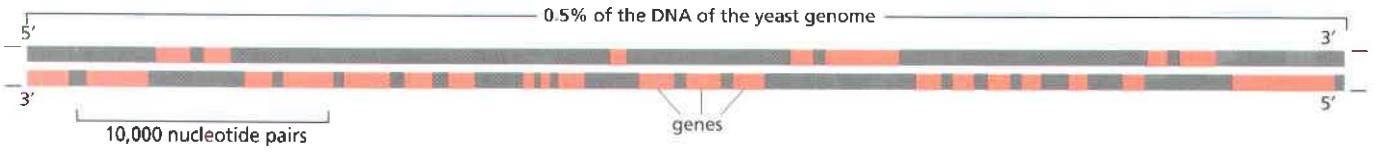


Figure 4–13 The arrangement of genes in the genome of *S. cerevisiae*. *S. cerevisiae* is a budding yeast widely used for brewing and baking. The genome of this yeast cell is distributed over 16 chromosomes. A small region of one chromosome has been arbitrarily selected to show the high density of genes characteristic of this species. As indicated by the *light red shading*, some genes are transcribed from the lower strand, while others are transcribed from the upper strand. There are about 6300 genes in the complete genome, which contains somewhat more than 12 million nucleotide pairs. (For the closely packed genes of a bacterium whose genome is 4.6 million nucleotides long, see Figure 1–29).

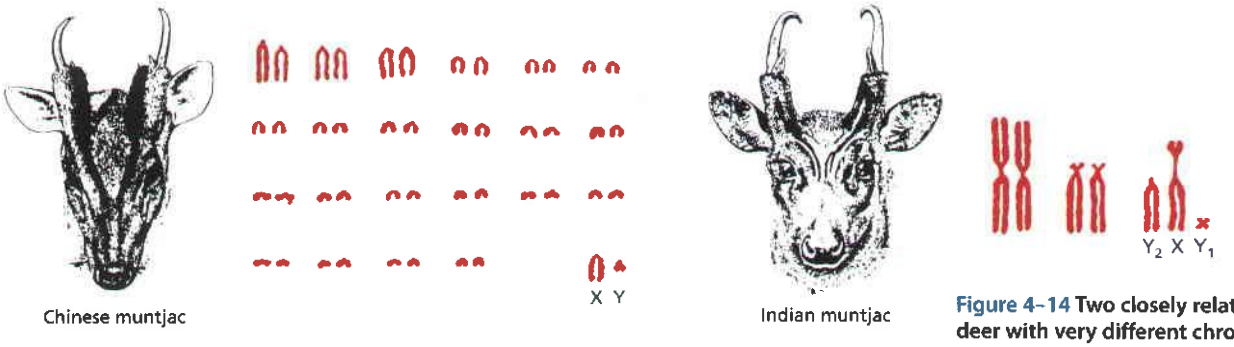


Figure 4–14 Two closely related species of deer with very different chromosome numbers. In the evolution of the Indian muntjac, initially separate chromosomes fused, without having a major effect on the animal. These two species contain a similar number of genes. (Adapted from M.W. Strickberger, Evolution, 3rd ed. Sudbury, MA: Jones & Bartlett Publishers, 2000.)

species complexity, and total genome size. Rather, the genomes and chromosomes of modern-day species have each been shaped by a unique history of seemingly random genetic events, acted on by selection pressures over long evolutionary times.

The Nucleotide Sequence of the Human Genome Shows How Our Genes Are Arranged

In Chapter 1 we discussed, in general terms, how the information in DNA is read out and used, through RNA intermediates, to make proteins (see Figure 1–4). In 1999, it became possible for the first time to see exactly how genes are arranged along an entire vertebrate chromosome (**Figure 4–15**). Today, with the publication of the "first draft" of the entire human genome in 2001 and the "finished DNA sequence" in 2004, the genetic information in all human chromosomes is available. The sheer quantity of information provided by the Human Genome Project is staggering (**Figure 4–16** and **Table 4–1**). At its peak, the Project generated raw nucleotide sequences at a rate of 1000 nucleotides per second around the clock. It will be many decades before this information is fully analyzed, but it has already stimulated new experiments that have had major effects on the content of every chapter in this book.

The first striking feature of the human genome is how little of it (only a few percent) codes for proteins (**Figure 4–17**). Much of the remaining chromosomal

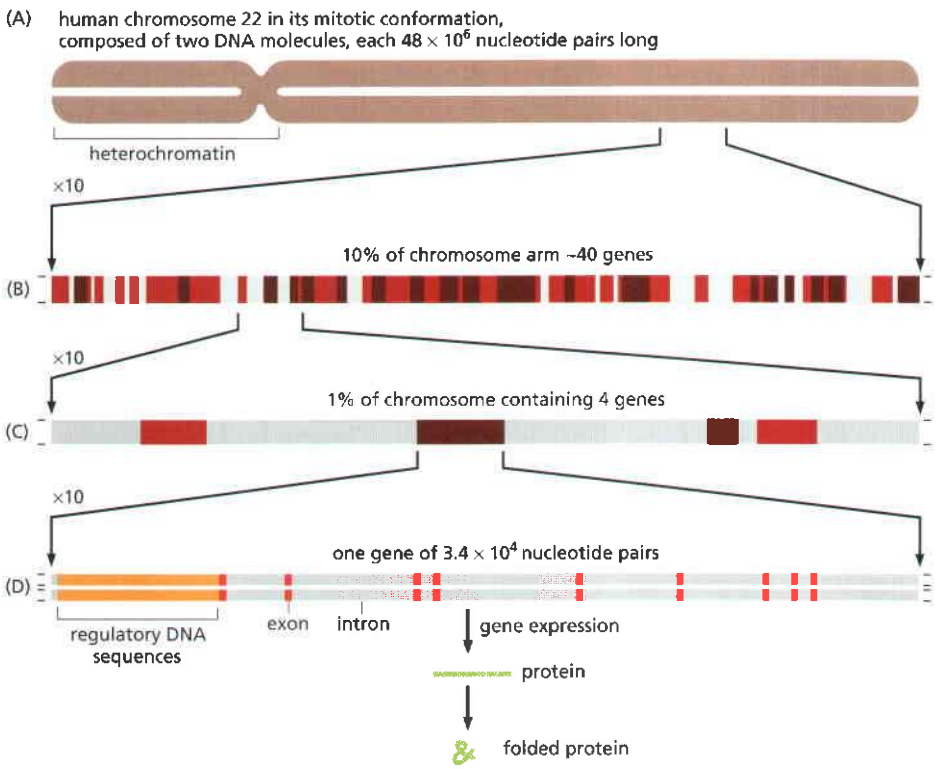


Figure 4–15 The organization of genes on a human chromosome. (A) Chromosome 22, one of the smallest human chromosomes, contains 48×10^6 nucleotide pairs and makes up approximately 1.5% of the entire human genome. Most of the left arm of chromosome 22 consists of short repeated sequences of DNA that are packaged in a particularly compact form of chromatin (heterochromatin), which is discussed later in this chapter. (B) A tenfold expansion of a portion of chromosome 22, with about 40 genes indicated. Those in *dark brown* are known genes and those in *red* are predicted genes. (C) An expanded portion of (B) shows the entire length of several genes. (D) The intron–exon arrangement of a typical gene is shown after a further tenfold expansion. Each exon *(red)* codes for a portion of the protein, while the DNA sequence of the introns *(gray)* is relatively unimportant, as discussed in detail in Chapter 6.

The human genome (3.2×10^9 nucleotide pairs) is the totality of genetic information belonging to our species. Almost all of this genome is distributed over the 22 autosomes and 2 sex chromosomes (see Figures 4–10 and 4–11) found within the nucleus. A minute fraction of the human genome (16,569 nucleotide pairs—in multiple copies per cell) is found in the mitochondria (introduced in Chapter 1, and discussed in detail in Chapter 14). The term *human genome sequence* refers to the complete nucleotide sequence of DNA in the 24 nuclear chromosomes and the mitochondria. Being diploid, a human somatic cell nucleus contains roughly twice the haploid amount of DNA, or 6.4×10^9 nucleotide pairs when not duplicating its chromosomes in preparation for division. (Adapted from International Human Genome Sequencing Consortium, *Nature* 409:860–921, 2001. With permission from Macmillan Publishers Ltd.)

DNA is made up of short, mobile pieces of DNA that have gradually inserted themselves in the chromosome over evolutionary time. We discuss these *transposable elements* in detail in later chapters.

A second notable feature of the human genome is the large average gene size of 27,000 nucleotide pairs. As discussed above, a typical gene carries in its linear sequence of nucleotides the information for the linear sequence of the amino acids of a protein. Only about 1300 nucleotide pairs are required to encode a protein of average size (about 430 amino acids in humans). Most of the remaining DNA in a gene consists of long stretches of noncoding DNA that interrupt the relatively short segments of DNA that code for protein. As will be discussed in detail in Chapter 6, the coding sequences are called **exons**; the intervening (noncoding) sequences in genes are called **introns** (see Figure 4–15 and Table 4–1). The majority of human genes thus consist of a long string of alternating exons and introns, with most of the gene consisting of introns. In contrast, the majority of genes from organisms with concise genomes lack introns. This accounts for the much smaller size of their genes (about one-twentieth that of human genes), as well as for the much higher fraction of coding DNA in their chromosomes.

In addition to introns and exons, each gene is associated with *regulatory DNA sequences*, which are responsible for ensuring that the gene is turned on or off at the proper time, expressed at the appropriate level, and only in the proper type of cell. In humans, the regulatory sequences for a typical gene are spread out over tens of thousands of nucleotide pairs. As would be expected, these regulatory sequences are more compressed in organisms with concise genomes. We discuss in Chapter 7 how regulatory DNA sequences work.

Finally, the nucleotide sequence of the human genome has revealed that the critical information needed to produce a human seems to be in an alarming state of disarray. As one commentator described our genome, "In some ways it may resemble your garage/bedroom/refrigerator/life: highly individualistic, but unkempt; little evidence of organization; much accumulated clutter (referred to by the uninitiated as 'junk'); virtually nothing ever discarded; and the few patently valuable items indiscriminately, apparently carelessly, scattered throughout."

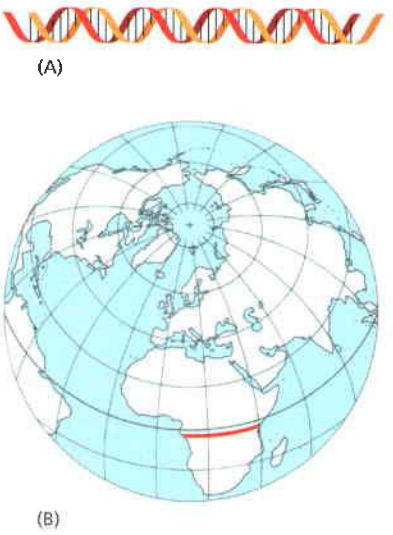


Figure 4–16 Scale of the human genome. If drawn with a 1 mm space between each nucleotide, as in (A), the human genome would extend 3200 km (approximately 2000 miles), far enough to stretch across the center of Africa, the site of our human origins *(red line* in B). At this scale, there would be, on average, a protein-coding gene every 130 m. An average gene would extend for 30 m, but the coding sequences in this gene would add up to only just over a meter.

Table 4–1 Some Vital Statistics for the Human Genome

	HUMAN GENOME
DNA length	3.2×10^9 nucleotide pairs*
Number of genes	approximately 25,000
Largest gene	2.4×10^6 nucleotide pairs
Mean gene size	27,000 nucleotide pairs
Smallest number of exons per gene	1
Largest number of exons per gene	178
Mean number of exons per gene	10.4
Largest exon size	17,106 nucleotide pairs
Mean exon size	145 nucleotide pairs
Number of pseudogenes**	more than 20,000
Percentage of DNA sequence in exons (protein coding sequences)	1.5%
Percentage of DNA in other highly conserved sequences***	3.5%
Percentage of DNA in high-copy repetitive elements	approximately 50%

* The sequence of 2.85 billion nucleotides is known precisely (error rate of only about one in 100,000 nucleotides). The remaining DNA primarily consists of short highly repeated sequences that are tandemly repeated, with repeat numbers differing from one individual to the next.

** A pseudogene is a nucleotide sequence of DNA closely resembling that of a functional gene, but containing numerous mutations that prevent its proper expression. Most pseudogenes arise from the duplication of a functional gene followed by the accumulation of damaging mutations in one copy.

*** Preserved functional regions; these include DNA encoding 5′ and 3′ UTRs (untranslated regions), structural and functional RNAs, and conserved protein-binding sites on the DNA.

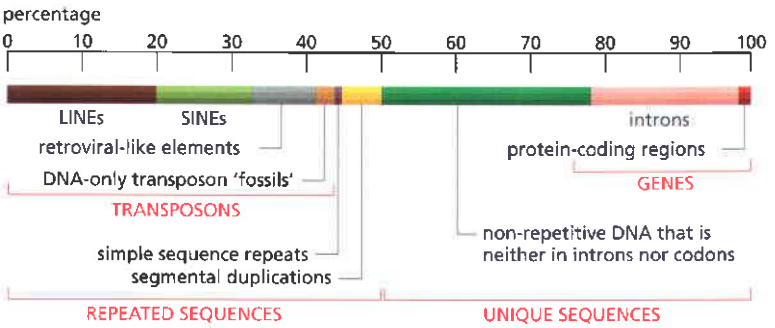


Figure 4–17 Representation of the nucleotide sequence content of the completely sequenced human genome. The LINEs, SINEs, retroviral-like elements, and DNA-only transposons are mobile genetic elements that have multiplied in our genome by replicating themselves and inserting the new copies in different positions. These mobile genetic elements are discussed in Chapter 5 (see Table 5–3, p. 318). Simple sequence repeats are short nucleotide sequences (less than 14 nucleotide pairs) that are repeated again and again for long stretches. Segmental duplications are large blocks of the genome (1000–200,000 nucleotide pairs) that are present at two or more locations in the genome. The most highly repeated blocks of DNA in heterochromatin have not yet been completely sequenced; therefore about 10% of human DNA sequences are not represented in this diagram. (Data courtesy of E. Margulies.)

Genome Comparisons Reveal Evolutionarily Conserved DNA Sequences

A major obstacle in interpreting the nucleotide sequences of human chromosomes is the fact that much of the sequence is probably unimportant. Moreover, the coding regions of the genome (the exons) are typically found in short segments (average size about 145 nucleotide pairs) floating in a sea of DNA whose exact nucleotide sequence is of little consequence. This arrangement makes it very difficult to identify all the exons in a stretch of DNA sequence. Even harder is the determination of where a gene begins and ends and exactly how many exons it spans.

Accurate gene identification requires approaches that extract information from the inherently low signal-to-noise ratio of the human genome. We shall describe some of them in Chapter 8. Here we discuss only one general approach, which is based on the observation that sequences that have a function are relatively conserved during evolution, whereas those without a function are free to mutate randomly. The strategy is therefore to compare the human sequence with that of the corresponding regions of a related genome, such as that of the mouse. Humans and mice are thought to have diverged from a common mammalian ancestor about 80×10^6 years ago, which is long enough for the majority of nucleotides in their genomes to have been changed by random mutational events. Consequently, the only regions that will have remained closely similar in the two genomes are those in which mutations would have impaired function and put the animals carrying them at a disadvantage, resulting in their elimination from the population by natural selection. Such closely similar regions are known as *conserved regions*. The conserved regions include both functionally important exons and regulatory DNA sequences. In contrast, *nonconserved regions* represent DNA whose sequence is unlikely to be critical for function.

The power of this method can be increased by comparing our genome with the genomes of additional animals whose genomes have been completely sequenced, including the rat, chicken, chimpanzee, and dog. By revealing in this way the results of a very long natural "experiment," lasting for hundreds of millions of years, such comparative DNA sequencing studies have highlighted the most interesting regions in these genomes. The comparisons reveal that roughly 5% of the human genome consists of "multi-species conserved sequences," as discussed in detail near the end of this chapter. Unexpectedly, only about one-third of these sequences code for proteins. Some of the conserved noncoding sequences correspond to clusters of protein-binding sites that are involved in gene regulation, while others produce RNA molecules that are not translated into protein. But the function of the majority of these sequences remains unknown. This unexpected discovery has led scientists to conclude that we understand much less about the cell biology of vertebrates than we had previously imagined. Certainly, there are enormous opportunities for new discoveries, and we should expect many surprises ahead.

Comparative studies have revealed not only that humans and other mammals share most of the same genes, but also that large blocks of our genomes contain these genes in the same order, a feature called *conserved synteny*. As a result, large blocks of our chromosomes can be recognized in other species. This allows the chromosome painting technique to be used to reconstruct the recent evolutionary history of human chromosomes (**Figure 4–18**).

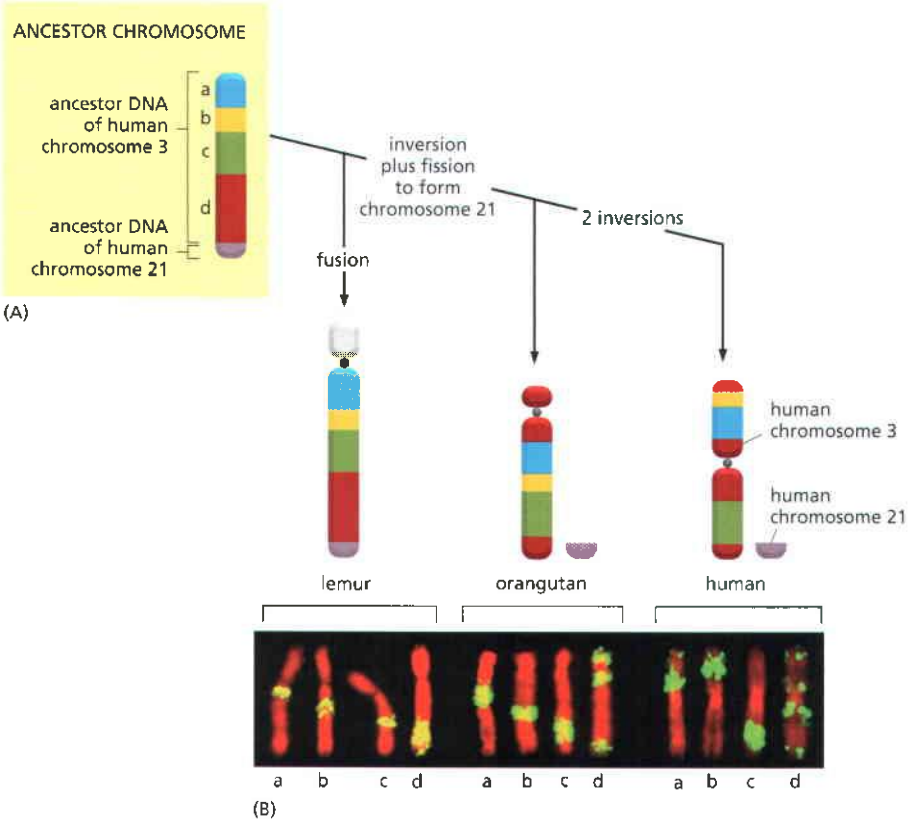


Figure 4–18 A proposed evolutionary history of human chromosome 3 and its relatives in other mammals. (A) The order of chromosome 3 segments hypothesized to be present on a chromosome of a mammalian ancestor is shown (*yellow* box). The minimum changes in this ancestral chromosome necessary to account for the appearance of each of the three modern chromosomes are indicated. (The present-day chromosomes of humans and African apes are identical at this resolution.) The *small circles* depicted in the modern chromosomes represent the positions of centromeres. A fission and inversion that leads to a change in chromosome organization is thought to occur once every $5–10 \times 10^6$ years in mammals. (B) Some of the chromosome painting experiments that led to the diagram in (A). Each image shows the chromosome most closely related to human chromosome 3, painted *green* by hybridization with different segments of DNA, lettered a, b, c, and d along the *bottom* of the figure. These letters correspond to the colored segments of the diagrams in (A), as indicated on the ancestral chromosome. (From S. Müller et al., *Proc. Natl Acad. Sci. U.S.A.* 97:206–211, 2000. With permission from National Academy of Sciences.)

Chromosomes Exist in Different States Throughout the Life of a Cell

We have seen how genes are arranged in chromosomes, but to form a functional chromosome, a DNA molecule must be able to do more than simply carry genes: it must be able to replicate, and the replicated copies must be separated and reliably partitioned into daughter cells at each cell division. This process occurs through an ordered series of stages, collectively known as the **cell cycle**, which provides for a temporal separation between the duplication of chromosomes and their segregation into two daughter cells. The cell cycle is briefly summarized in **Figure 4–19**, and it is discussed in detail in Chapter 17. Only certain

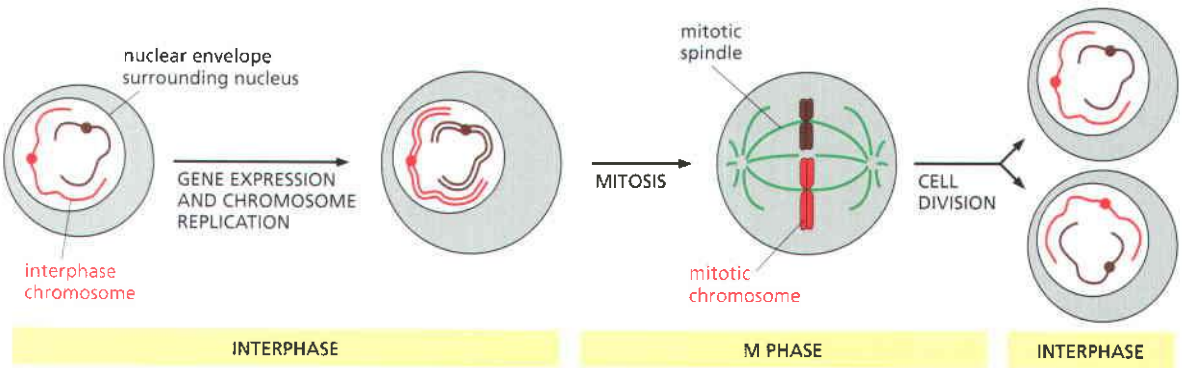


Figure 4–19 A simplified view of the eucaryotic cell cycle. During interphase, the cell is actively expressing its genes and is therefore synthesizing proteins. Also, during interphase and before cell division, the DNA is replicated and each chromosome is duplicated to produce two closely paired daughter chromosomes (a cell with only two chromosomes is illustrated here). Once DNA replication is complete, the cell can enter *M phase*, when mitosis occurs and the nucleus is divided into two daughter nuclei. During this stage, the chromosomes condense, the nuclear envelope breaks down, and the mitotic spindle forms from microtubules and other proteins. The condensed mitotic chromosomes are captured by the mitotic spindle, and one complete set of chromosomes is then pulled to each end of the cell by separating each daughter chromosome pair. A nuclear envelope re-forms around each chromosome set, and in the final step of M phase, the cell divides to produce two daughter cells. Most of the time in the cell cycle is spent in interphase; M phase is brief in comparison, occupying only about an hour in many mammalian cells.

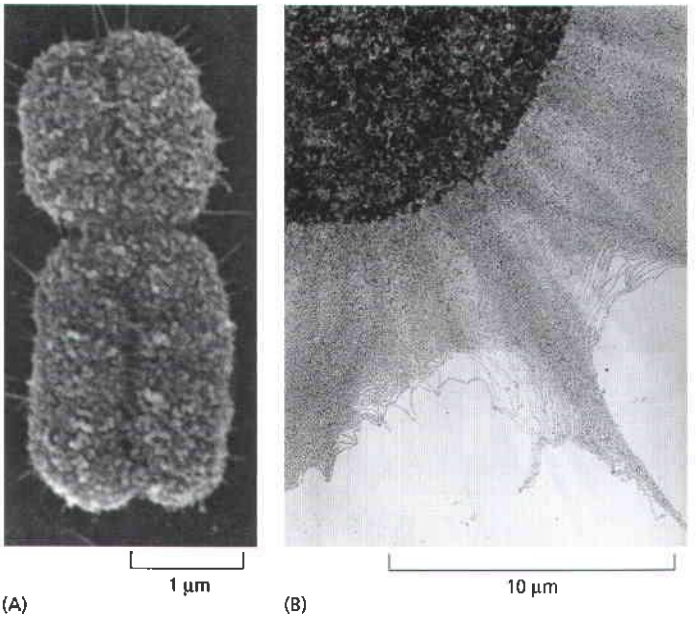


Figure 4–20 A comparison of extended interphase chromatin with the chromatin in a mitotic chromosome. (A) A scanning electron micrograph of a mitotic chromosome: a condensed duplicated chromosome in which the two new chromosomes are still linked together (see Figure 4–21). The constricted region indicates the position of the centromere, described in Figure 4–21. (B) An electron micrograph showing an enormous tangle of chromatin spilling out of a lysed interphase nucleus. Note the difference in scales. (A, courtesy of Terry D. Allen; B, courtesy of Victoria Foe.)

parts of the cycle concern us in this chapter. During *interphase* chromosomes are replicated, and during *mitosis* they become highly condensed and then are separated and distributed to the two daughter nuclei. The highly condensed chromosomes in a dividing cell are known as *mitotic chromosomes* (**Figure 4–20**A). This is the form in which chromosomes are most easily visualized; in fact, the images of chromosomes shown so far in the chapter are of chromosomes in mitosis. During cell division, this condensed state is important for the accurate separation of the duplicated chromosomes by the mitotic spindle, as discussed in Chapter 17.

During the portions of the cell cycle when the cell is not dividing, the chromosomes are extended and much of their chromatin exists as long, thin tangled threads in the nucleus so that individual chromosomes cannot be easily distinguished (Figure 4–20B). We shall refer to chromosomes in this extended state as *interphase chromosomes*. Since cells spend most of their time in interphase, and this is where their genetic information is being read out, chromosomes are of greatest interest to cell biologists when they are least visible.

Each DNA Molecule That Forms a Linear Chromosome Must Contain a Centromere, Two Telomeres, and Replication Origins

A chromosome operates as a distinct structural unit: for a copy to be passed on to each daughter cell at division, each chromosome must be able to replicate, and the newly replicated copies must subsequently be separated and partitioned correctly into the two daughter cells. These basic functions are controlled by three types of specialized nucleotide sequences in the DNA, each of which binds specific proteins that guide the machinery that replicates and segregates chromosomes (**Figure 4–21**).

Experiments in yeasts, whose chromosomes are relatively small and easy to manipulate, have identified the minimal DNA sequence elements responsible for each of these functions. One type of nucleotide sequence acts as a DNA **replication origin**, the location at which duplication of the DNA begins. Eucaryotic chromosomes contain many origins of replication to ensure that the entire chromosome can be replicated rapidly, as discussed in detail in Chapter 5.

After replication, the two daughter chromosomes remain attached to one another and, as the cell cycle proceeds, are condensed further to produce mitotic chromosomes. The presence of a second specialized DNA sequence, called a **centromere**, allows one copy of each duplicated and condensed chromosome to be pulled into each daughter cell when a cell divides. A protein

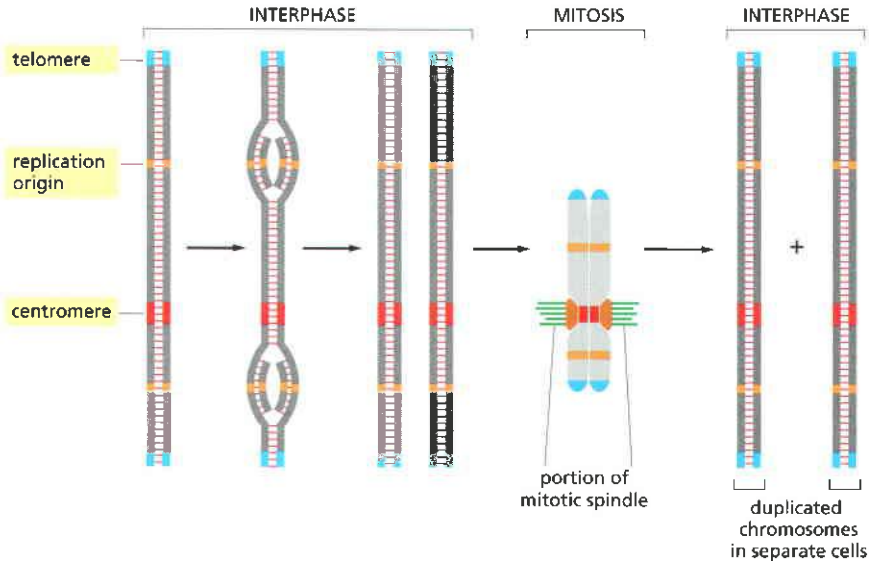


Figure 4–21 The three DNA sequences required to produce a eucaryotic chromosome that can be replicated and then segregated at mitosis. Each chromosome has multiple origins of replication, one centromere, and two telomeres. Shown here is the sequence of events that a typical chromosome follows during the cell cycle. The DNA replicates in interphase, beginning at the origins of replication and proceeding bidirectionally from the origins across the chromosome. In M phase, the centromere attaches the duplicated chromosomes to the mitotic spindle so that one copy is distributed to each daughter cell during mitosis. The centromere also helps to hold the duplicated chromosomes together until they are ready to be moved apart. The telomeres form special caps at each chromosome end.

complex called a *kinetochore* forms at the centromere and attaches the duplicated chromosomes to the mitotic spindle, allowing them to be pulled apart (discussed in Chapter 17).

The third specialized DNA sequence forms **telomeres**, the ends of a chromosome. Telomeres contain repeated nucleotide sequences that enable the ends of chromosomes to be efficiently replicated. Telomeres also perform another function: the repeated telomere DNA sequences, together with the regions adjoining them, form structures that protect the end of the chromosome from being mistaken by the cell for a broken DNA molecule in need of repair. We discuss both this type of repair and the structure and function of telomeres in Chapter 5.

In yeast cells, the three types of sequences required to propagate a chromosome are relatively short (typically less than 1000 base pairs each) and therefore use only a tiny fraction of the information-carrying capacity of a chromosome. Although telomere sequences are fairly simple and short in all eucaryotes, the DNA sequences that form centromeres and replication origins in more complex organisms are much longer than their yeast counterparts. For example, experiments suggest that human centromeres contain up to 100,000 nucleotide pairs and may not require a stretch of DNA with a defined nucleotide sequence. Instead, as we shall discuss later in this chapter, they seem to consist of a large, regularly repeating protein–nucleic acid structure that can be inherited when a chromosome replicates.

DNA Molecules Are Highly Condensed in Chromosomes

All eucaryotic organisms have special ways of packaging DNA into chromosomes. For example, if the 48 million nucleotide pairs of DNA in human chromosome 22 could be laid out as one long perfect double helix, the molecule would extend for about 1.5 cm if stretched out end to end. But chromosome 22 measures only about 2 μm in length in mitosis (see Figures 4–10 and 4–11), representing an end-to-end compaction ratio of nearly 10,000-fold. This remarkable feat of compression is performed by proteins that successively coil and fold the DNA into higher and higher levels of organization. Although much less condensed than mitotic chromosomes, the DNA of human interphase chromosomes is still tightly packed, with an overall compaction ratio of approximately 500-fold (the length of a chromosome's DNA helix divided by the end-to-end length of that chromosome).

In reading these sections it is important to keep in mind that chromosome structure is dynamic. We have seen that each chromosome condenses to an unusual degree in the M phase of the cell cycle. Much less visible, but of enormous interest and importance, specific regions of interphase chromosomes

decondense as the cells gain access to specific DNA sequences for gene expression, DNA repair, and replication—and then recondense when these processes are completed. The packaging of chromosomes is therefore accomplished in a way that allows rapid localized, on-demand access to the DNA. In the next sections we discuss the specialized proteins that make this type of packaging possible.

Nucleosomes Are a Basic Unit of Eucaryotic Chromosome Structure

The proteins that bind to the DNA to form eucaryotic chromosomes are traditionally divided into two general classes: the **histones** and the *nonhistone chromosomal proteins*. The complex of both classes of protein with the nuclear DNA of eucaryotic cells is known as **chromatin**. Histones are present in such enormous quantities in the cell (about 60 million molecules of each type per human cell) that their total mass in chromatin is about equal to that of the DNA.

Histones are responsible for the first and most basic level of chromosome packing, the **nucleosome**, a protein–DNA complex discovered in 1974. When interphase nuclei are broken open very gently and their contents examined under the electron microscope, most of the chromatin is in the form of a fiber with a diameter of about 30 nm (**Figure 4–22**A). If this chromatin is subjected to treatments that cause it to unfold partially, it can be seen under the electron microscope as a series of "beads on a string" (Figure 4–22B). The string is DNA, and each bead is a "nucleosome core particle" that consists of DNA wound around a protein core formed from histones. <ACTC>

The structural organization of nucleosomes was determined after first isolating them from unfolded chromatin by digestion with particular enzymes (called nucleases) that break down DNA by cutting between the nucleosomes. After digestion for a short period, the exposed DNA between the nucleosome core particles, the *linker DNA*, is degraded. Each individual nucleosome core particle consists of a complex of eight histone proteins—two molecules each of histones H2A, H2B, H3, and H4—and double-stranded DNA that is 147 nucleotide pairs long. The *histone octamer* forms a protein core around which the double-stranded DNA is wound (**Figure 4–23**).

Each nucleosome core particle is separated from the next by a region of linker DNA, which can vary in length from a few nucleotide pairs up to about 80. (The term *nucleosome* technically refers to a nucleosome core particle plus one of its adjacent DNA linkers, but it is often used synonymously with nucleosome core particle.) On average, therefore, nucleosomes repeat at intervals of about 200 nucleotide pairs. For example, a diploid human cell with 6.4×10^9 nucleotide pairs contains approximately 30 million nucleosomes. The formation of nucleosomes converts a DNA molecule into a chromatin thread about one-third of its initial length.

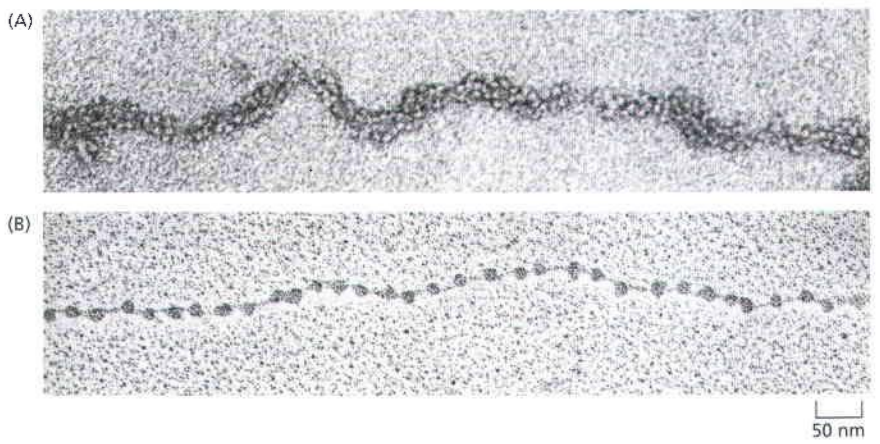


Figure 4–22 Nucleosomes as seen in the electron microscope. (A) Chromatin isolated directly from an interphase nucleus appears in the electron microscope as a thread 30 nm thick. (B) This electron micrograph shows a length of chromatin that has been experimentally unpacked, or decondensed, after isolation to show the nucleosomes. (A, courtesy of Barbara Hamkalo; B, courtesy of Victoria Foe.)

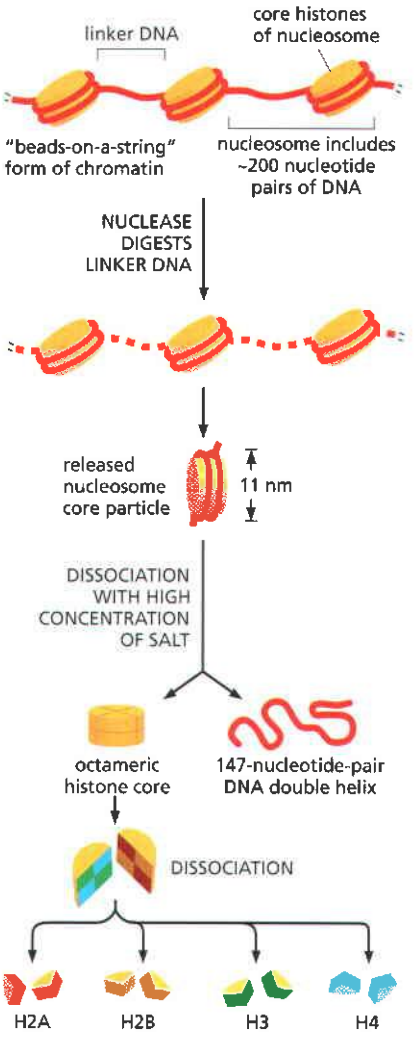


Figure 4–23 Structural organization of the nucleosome. A nucleosome contains a protein core made of eight histone molecules. In biochemical experiments, the nucleosome core particle can be released from isolated chromatin by digestion of the linker DNA with a nuclease, an enzyme that breaks down DNA. (The nuclease can degrade the exposed linker DNA but cannot attack the DNA wound tightly around the nucleosome core.) After dissociation of the isolated nucleosome into its protein core and DNA, the length of the DNA that was wound around the core can be determined. This length of 147 nucleotide pairs is sufficient to wrap 1.7 times around the histone core.

The Structure of the Nucleosome Core Particle Reveals How DNA Is Packaged

The high-resolution structure of a nucleosome core particle, solved in 1997, revealed a disc-shaped histone core around which the DNA was tightly wrapped 1.7 turns in a left-handed coil (Figure 4–24). All four of the histones that make up the core of the nucleosome are relatively small proteins (102–135 amino acids), and they share a structural motif, known as the *histone fold*, formed from three α helices connected by two loops (Figure 4–25). In assembling a nucleosome, the histone folds first bind to each other to form H3–H4 and H2A–H2B dimers, and the H3–H4 dimers combine to form tetramers. An H3–H4 tetramer then further combines with two H2A–H2B dimers to form the compact octamer core, around which the DNA is wound (Figure 4–26).

The interface between DNA and histone is extensive: 142 hydrogen bonds are formed between DNA and the histone core in each nucleosome. Nearly half of these bonds form between the amino acid backbone of the histones and the phosphodiester backbone of the DNA. Numerous hydrophobic interactions and salt linkages also hold DNA and protein together in the nucleosome. For example, more than one-fifth of the amino acids in each of the core histones are either lysine or arginine (two amino acids with basic side chains), and their positive

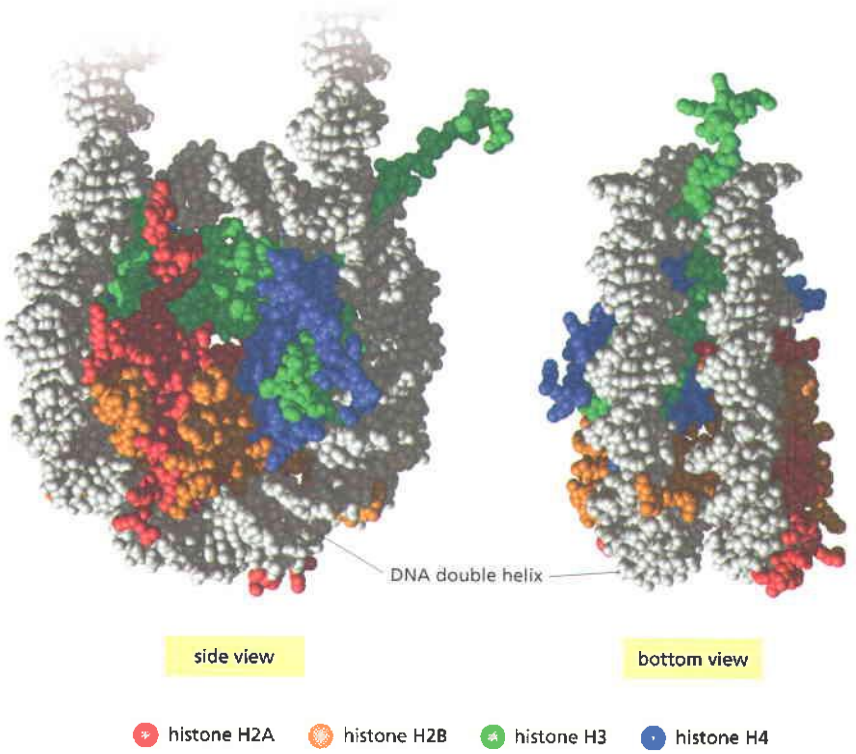


Figure 4–24 The structure of a nucleosome core particle, as determined by x-ray diffraction analyses of crystals. Each histone is colored according to the scheme in Figure 4–23, with the DNA double helix in *light gray*. (From K. Luger et al., *Nature* 389:251–260, 1997. With permission from Macmillan Publishers Ltd.)

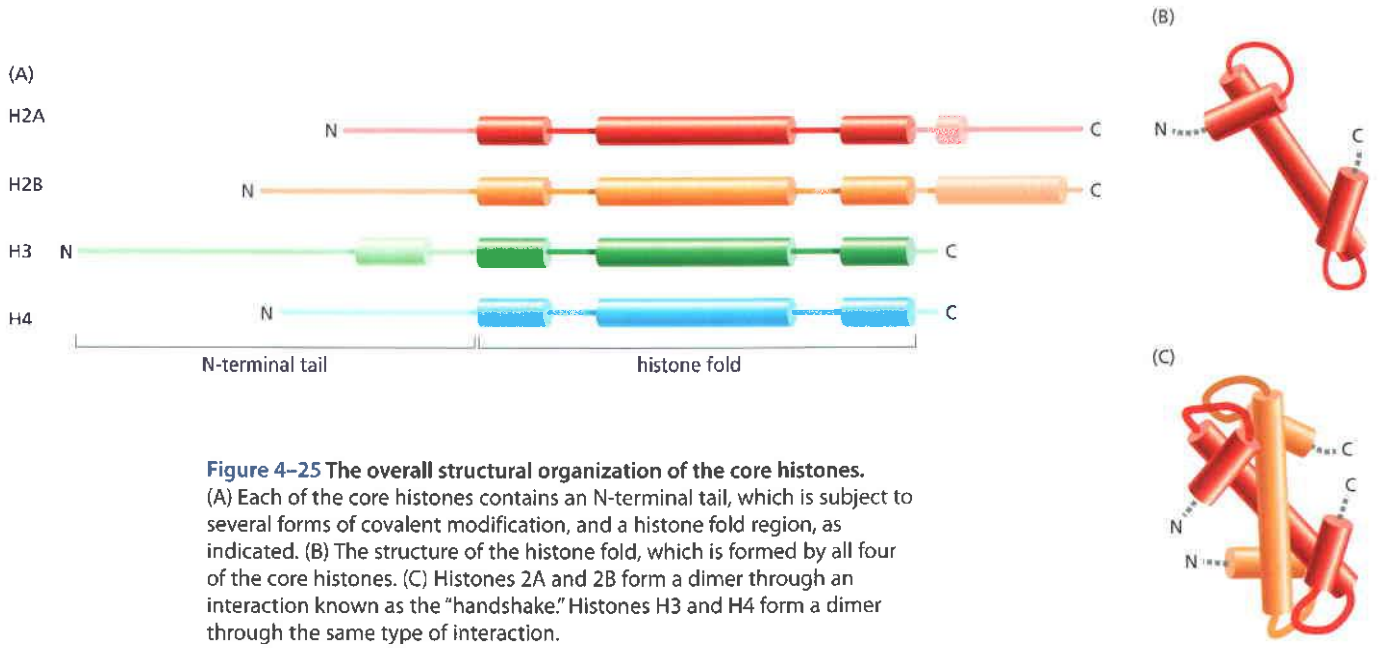


Figure 4–25 The overall structural organization of the core histones. (A) Each of the core histones contains an N-terminal tail, which is subject to several forms of covalent modification, and a histone fold region, as indicated. (B) The structure of the histone fold, which is formed by all four of the core histones. (C) Histones 2A and 2B form a dimer through an interaction known as the "handshake." Histones H3 and H4 form a dimer through the same type of interaction.

charges can effectively neutralize the negatively charged DNA backbone. These numerous interactions explain in part why DNA of virtually any sequence can be bound on a histone octamer core. The path of the DNA around the histone core is not smooth; rather, several kinks are seen in the DNA, as expected from the nonuniform surface of the core. The bending requires a substantial compression of the minor groove of the DNA helix. Certain dinucleotides in the minor groove are especially easy to compress, and some nucleotide sequences bind the nucleosome more tightly than others (**Figure 4–27**). This probably explains some striking, but unusual, cases of very precise positioning of nucleosomes along a stretch of DNA. For most of the DNA sequences found in chromosomes, however, the sequence preference of nucleosomes must be small enough to allow other factors to dominate, inasmuch as nucleosomes can occupy any one of a number of positions relative to the DNA sequence in most chromosomal regions.

In addition to its histone fold, each of the core histones has an N-terminal amino acid "tail", which extends out from the DNA–histone core (see Figure 4–26). These histone tails are subject to several different types of covalent modifications that in turn control critical aspects of chromatin structure and function, as we shall discuss shortly.

As a reflection of their fundamental role in DNA function through controlling chromatin structure, the histones are among the most highly conserved eucaryotic proteins. For example, the amino acid sequence of histone H4 from a pea and from a cow differ at only 2 of the 102 positions. This strong evolutionary conservation suggests that the functions of histones involve nearly all of their amino acids, so that a change in any position is deleterious to the cell. This suggestion has been tested directly in yeast cells, in which it is possible to mutate a given histone gene *in vitro* and introduce it into the yeast genome in place of the normal gene. As might be expected, most changes in histone sequences are lethal; the few that are not lethal cause changes in the normal pattern of gene expression, as well as other abnormalities.

Despite the high conservation of the core histones, eucaryotic organisms also produce smaller amounts of specialized variant core histones that differ in amino acid sequence from the main ones. As we shall see, these variants, combined with a surprisingly large variety of covalent modifications that can be added to the histones in nucleosomes, make possible the many different chromatin structures that are required for DNA function in higher eucaryotes.

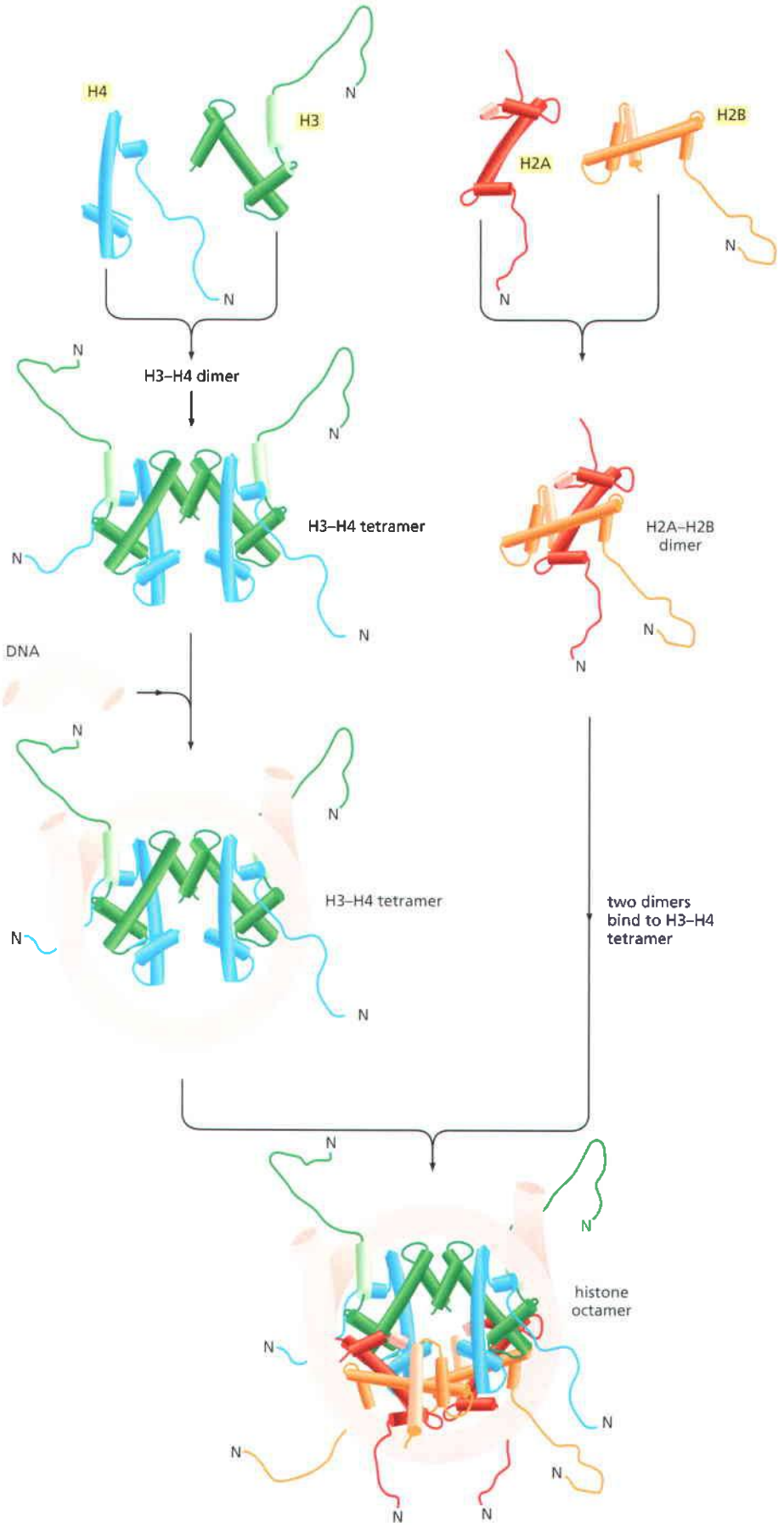


Figure 4–26 The assembly of a histone octamer on DNA. The histone H3–H4 dimer and the H2A–H2B dimer are formed from the handshake interaction. An H3–H4 tetramer forms and binds to the DNA. Two H2A–H2B dimers are then added, to complete the nucleosome. The histones are colored as in Figures 4–24 and 4–25. Note that all eight N-terminal tails of the histones protrude from the disc-shaped core structure. Their conformations are highly flexible.

Inside the cell, the nucleosome assembly reactions shown here are mediated by *histone chaperone* proteins, some specific for H3–H4 and others specific for H2A–H2B. (Adapted from figures by J. Waterborg.)

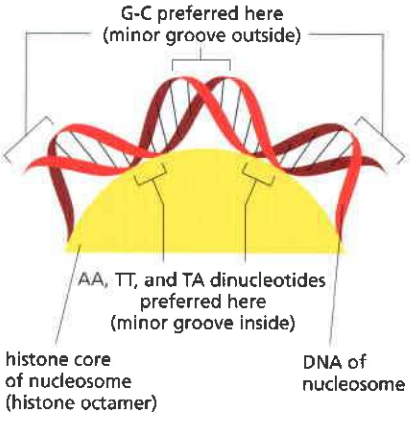


Figure 4–27 The bending of DNA in a nucleosome. The DNA helix makes 1.7 tight turns around the histone octamer. This diagram illustrates how the minor groove is compressed on the inside of the turn. Owing to certain structural features of the DNA molecule, the indicated dinucleotides are preferentially accommodated in such a narrow minor groove, which helps to explain why certain DNA sequences will bind more tightly than others to the nucleosome core.

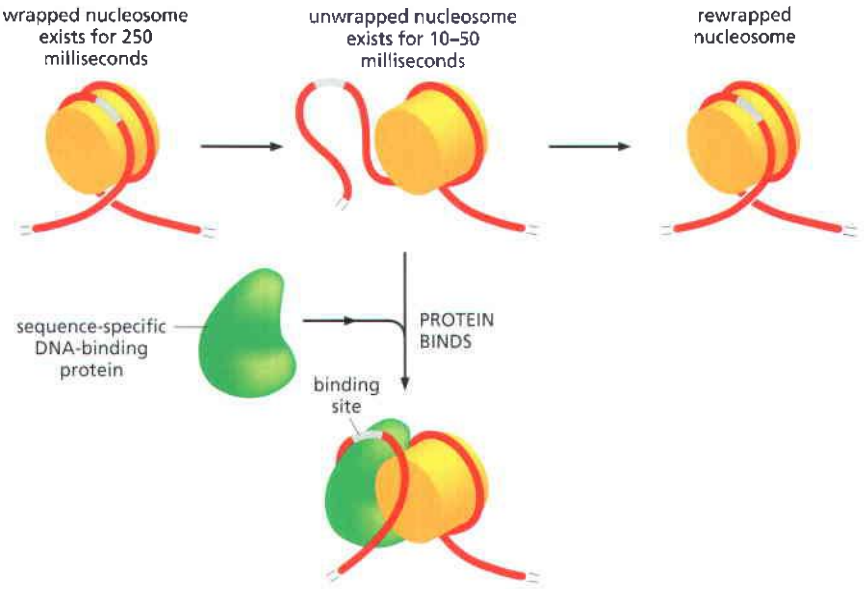


Figure 4–28 Dynamic nucleosomes. Kinetic measurements show that the DNA in an isolated nucleosome is surprisingly dynamic, rapidly uncoiling and then rewrapping around its nucleosome core. As indicated, this makes most of its bound DNA sequence accessible to other DNA-binding proteins. (Data from G. Li and J. Widom, *Nat. Struct. Mol. Biol.* 11:763–769, 2004. With permission from Macmillan Publishers Ltd.)

Nucleosomes Have a Dynamic Structure, and Are Frequently Subjected to Changes Catalyzed by ATP-Dependent Chromatin-Remodeling Complexes

For many years biologists thought that, once formed in a particular position on DNA, a nucleosome remains fixed in place because of the very tight association between its core histones and DNA. If true, this would pose problems for genetic readout mechanisms, which in principle require rapid access to many specific DNA sequences, as well as for the rapid passage of the DNA transcription and replication machinery through chromatin. But kinetic experiments show that the DNA in an isolated nucleosome unwraps from each end at rate of about 4 times per second, remaining exposed for 10 to 50 milliseconds before the partially unwrapped structure recloses. Thus, most of the DNA in an isolated nucleosome is in principle available for binding other proteins (Figure 4–28).

For the chromatin in a cell, a further loosening of DNA–histone contacts is clearly required, because eucaryotic cells contain a large variety of ATP-dependent *chromatin remodeling complexes*. The subunit in these complexes that hydrolyzes ATP is evolutionarily related to the DNA helicases (discussed in Chapter 5), and it binds both to the protein core of the nucleosome and to the double-stranded DNA that winds around it. By using the energy of ATP hydrolysis to move this DNA relative to the core, this subunit changes the structure of a nucleosome temporarily, making the DNA less tightly bound to the histone core. Through repeated cycles of ATP hydrolysis, the remodeling complexes can catalyze *nucleosome sliding*, and by pulling the nucleosome core along the DNA double helix in this way, they make the nucleosomal DNA available to other proteins in the cell (Figure 4–29). In addition, by cooperating with negatively

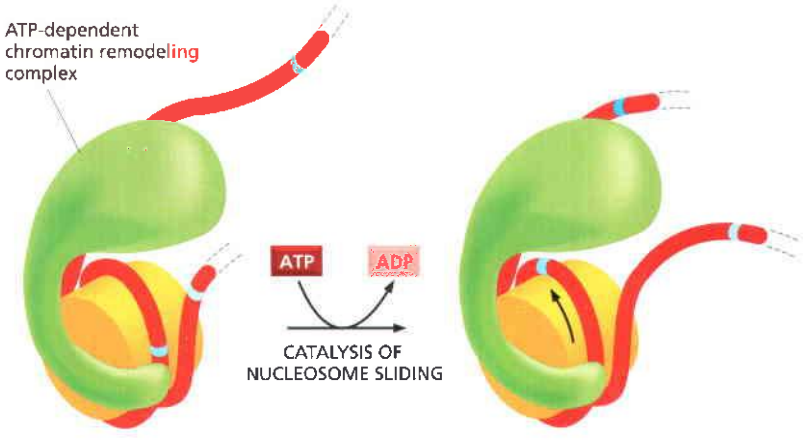


Figure 4–29 The nucleosome sliding catalyzed by ATP-dependent chromatin remodeling complexes. Using the energy of ATP hydrolysis, the remodeling complex is thought to push on the DNA of its bound nucleosome and loosen its attachment to the nucleosome core. Each cycle of ATP binding, ATP hydrolysis, and release of the ADP and P_i products thereby moves the DNA with respect to the histone octamer in the direction of the arrow in this diagram. It requires many such cycles to produce the nucleosome sliding shown. (See also Figure 4–46B.)

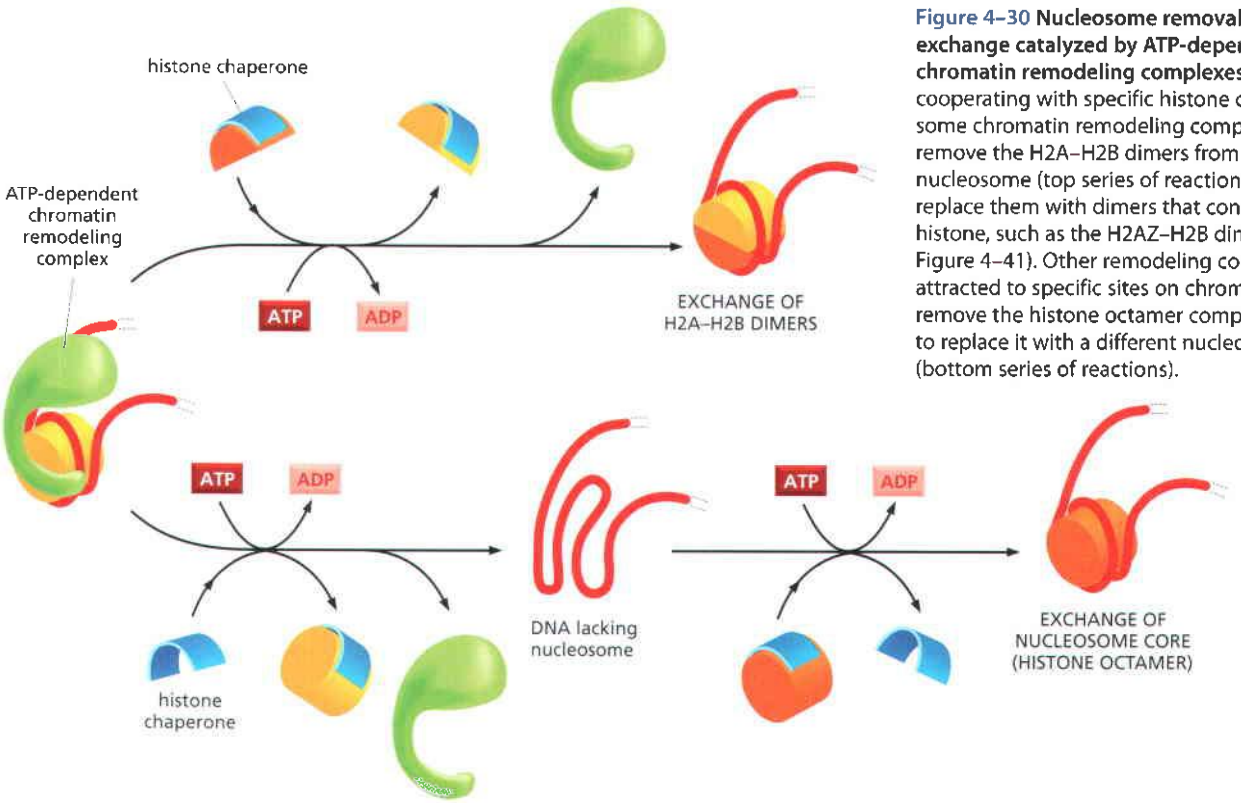


Figure 4–30 Nucleosome removal and histone exchange catalyzed by ATP-dependent chromatin remodeling complexes. By cooperating with specific histone chaperones, some chromatin remodeling complexes can remove the H2A–H2B dimers from a nucleosome (top series of reactions) and replace them with dimers that contain a variant histone, such as the H2AZ–H2B dimer (see Figure 4–41). Other remodeling complexes are attracted to specific sites on chromatin to remove the histone octamer completely and/or to replace it with a different nucleosome core (bottom series of reactions).

charged proteins that serve as histone chaperones, some remodeling complexes are able to remove either all or part of the nucleosome core from a nucleosome—catalyzing either an exchange of its H2A–H2B histones, or the complete removal of the octameric core from the DNA (**Figure 4–30**).

Cells contain dozens of different ATP-dependent chromatin remodeling complexes that are specialized for different roles. Most are large protein complexes that can contain 10 or more subunits. The activity of these complexes is carefully controlled by the cell. As genes are turned on and off, chromatin remodeling complexes are brought to specific regions of DNA where they act locally to influence chromatin structure (discussed in Chapter 7; see also Figure 4–46, below).

As pointed out previously, for most of the DNA sequences found in chromosomes, experiments show that a nucleosome can occupy any one of a number of positions relative to the DNA sequence. The most important influence on nucleosome positioning appears to be the presence of other tightly bound proteins on the DNA. Some bound proteins favor the formation of a nucleosome adjacent to them. Others create obstacles that force the nucleosomes to move to positions between them. The exact positions of nucleosomes along a stretch of DNA therefore depends mainly on the presence and nature of other proteins bound to the DNA. Due to the presence of ATP-dependent remodeling complexes, the arrangement of nucleosomes on DNA can be highly dynamic, changing rapidly according to the needs of the cell.

Nucleosomes Are Usually Packed Together into a Compact Chromatin Fiber

Although enormously long strings of nucleosomes form on the chromosomal DNA, chromatin in a living cell probably rarely adopts the extended "beads on a string" form. Instead, the nucleosomes are packed on top of one another, generating regular arrays in which the DNA is even more highly condensed. Thus, when nuclei are very gently lysed onto an electron microscope grid, most of the chromatin is seen to be in the form of a fiber with a diameter of about 30 nm, which is considerably wider than chromatin in the "beads on a string" form (see Figure 4–22).

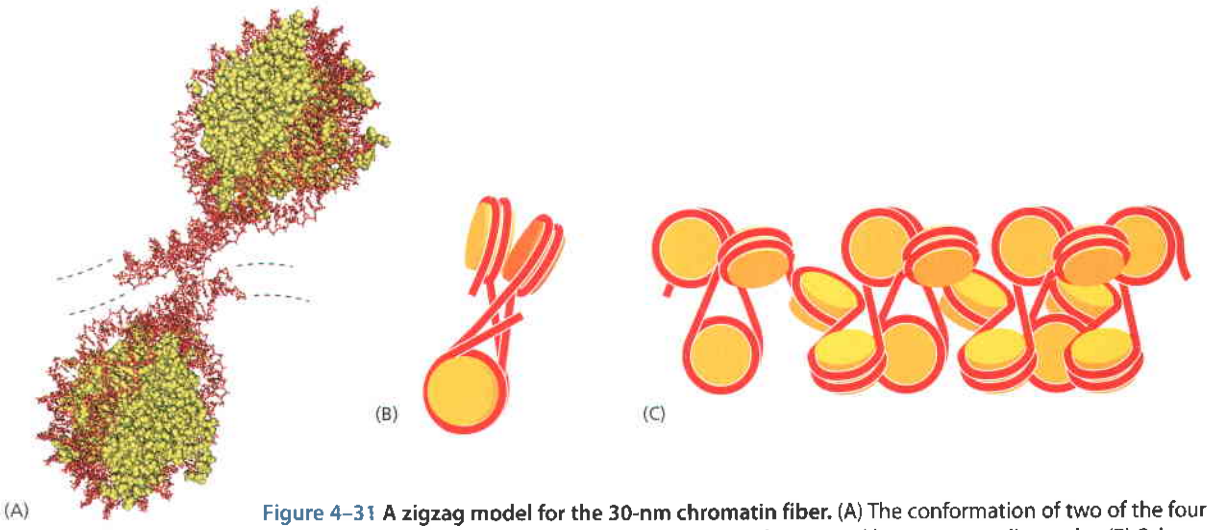


Figure 4–31 A zigzag model for the 30-nm chromatin fiber. (A) The conformation of two of the four nucleosomes in a tetranucleosome, from a structure determined by x-ray crystallography. (B) Schematic of the entire tetranucleosome; the fourth nucleosome is not visible, being stacked on the bottom nucleosome and behind it in this diagram. (C) Diagrammatic illustration of a possible zigzag structure that could account for the 30-nm chromatin fiber. (Adapted from C.L. Woodcock, *Nat. Struct. Mol. Biol.* 12:639–640, 2005. With permission from Macmillan Publishers Ltd.)

How are nucleosomes packed in the 30-nm chromatin fiber? This question has not yet been answered definitively, but important information concerning the structure has been obtained. In particular, high-resolution structural analyses have been performed on homogeneous short strings of nucleosomes, prepared from purified histones and purified DNA molecules. The structure of a tetranucleosome, obtained by X-ray crystallography, has been used to support a zigzag model for the stacking of nucleosomes in the 30-nm fiber (**Figure 4–31**). But cryoelectron microscopy of longer strings of nucleosomes supports a very different solenoidal structure with intercalated nucleosomes (**Figure 4–32**).

What causes the nucleosomes to stack so tightly on each other in a 30-nm fiber? The nucleosome to nucleosome linkages formed by histone tails, most notably the H4 tail (**Figure 4–33**) constitute one important factor. Another

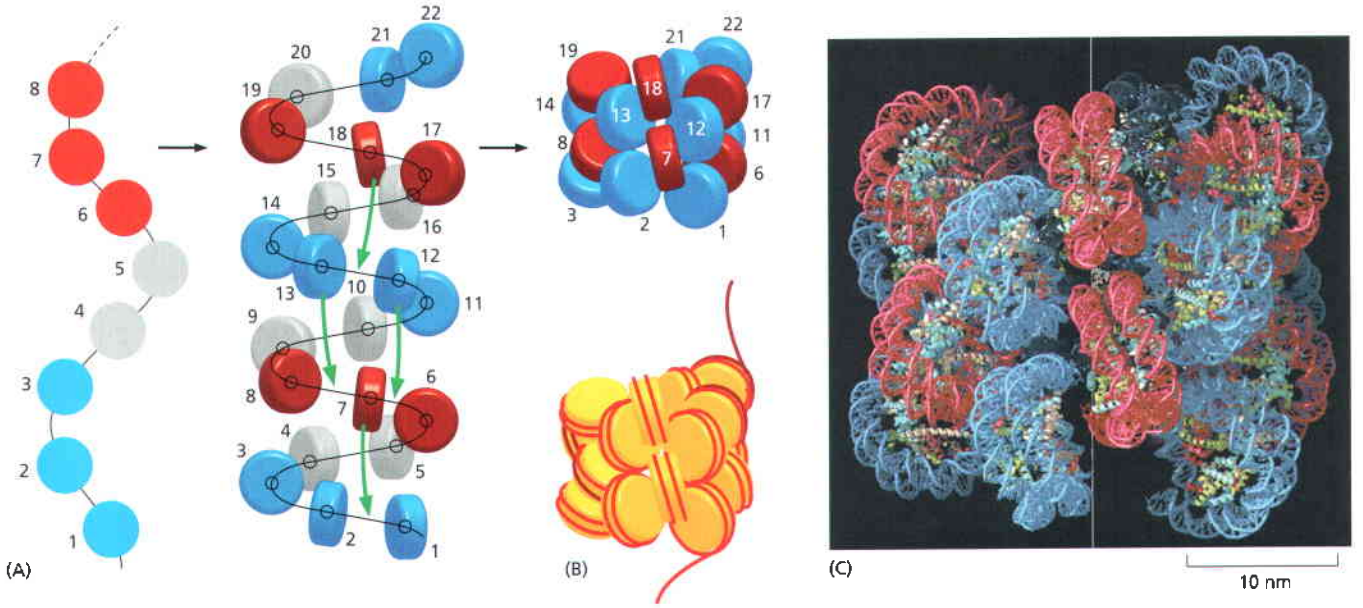


Figure 4–32 An interdigitated solenoid model for the 30-nm chromatin fiber. (A) Drawings in which strings of color-coded nucleosomes are used to illustrate how the solenoid is generated. (B) Schematic diagram of final structure in (A). (C) Structural model. The model is derived from high-resolution cryoelectron microscopy images of nucleosome arrays reconstituted from purified histones and DNA molecules of specific length and sequence. Both nucleosome octamers and a linker histone (discussed below) were used to produce regularly repeating arrays containing up to 72 nucleosomes. (Adapted from P. Robinson, L. Fairall, V. Huynh and D. Rhodes, *Proc. Natl Acad. Sci. U.S.A.* 103:6506–6511, 2006. With permission from National Academy of Sciences.)

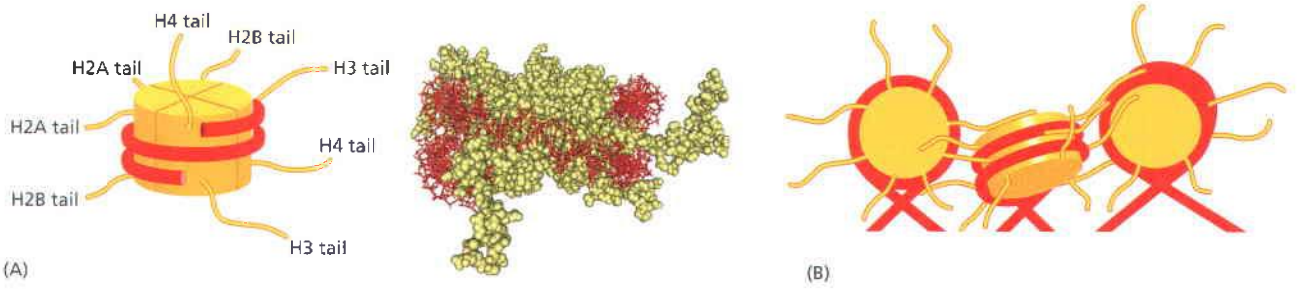


Figure 4–33 A speculative model for the role played by histone tails in the formation of the 30-nm fiber. (A) This schematic diagram shows the approximate exit points of the eight histone tails, one from each histone protein, that extend from each nucleosome. The actual structure is shown to its right. In the high-resolution structure of the nucleosome, the tails are largely unstructured, suggesting that they are highly flexible. (B) A speculative model showing how the histone tails may help to pack nucleosomes together into the 30-nm fiber. This model is based on (1) experimental evidence that histone tails aid in the formation of the 30-nm fiber, and (2) the x-ray crystal structure of the nucleosome, in which the tails of one nucleosome contact the histone core of an adjacent nucleosome in the crystal lattice.

important factor is an additional histone that is often present in a 1-to-1 ratio with nucleosome cores, known as **histone H1**. This so-called linker histone is larger than the individual core histones and it has been considerably less well conserved during evolution. A single histone H1 molecule binds to each nucleosome, contacting both DNA and protein, and changing the path of the DNA as it exits from the nucleosome. Although it is not understood in detail how H1 pulls nucleosomes together into the 30-nm fiber, a change in the exit path in DNA seems crucial for compacting nucleosomal DNA so that it interlocks to form the 30-nm fiber (**Figure 4–34**). Most eucaryotic organisms make several histone H1 proteins of related but quite distinct amino acid sequences.

It is possible that the 30-nm structure found in chromosomes is a fluid mosaic of several different variations. For example, a linker histone in the H1 family was present in the nucleosomal arrays studied in Figure 4–32 but was missing from the tetranucleosome in Figure 4–31. Moreover, we saw earlier that the linker DNA that connects adjacent nucleosomes can vary in length; these differences in linker length probably introduce local perturbations into the structure. And the presence of many other DNA-binding proteins, as well as proteins that bind directly to histones, will certainly add important additional features to any array of nucleosomes.

Summary

A gene is a nucleotide sequence in a DNA molecule that acts as a functional unit for the production of a protein, a structural RNA, or a catalytic or regulatory RNA molecule. In eucaryotes, protein-coding genes are usually composed of a string of alternating introns and exons associated with regulatory regions of DNA. A chromosome is formed from a single, enormously long DNA molecule that contains a linear array of many genes. The human genome contains 3.2 × 10^9 DNA nucleotide pairs, divided between 22 different autosomes and 2 sex chromosomes. Only a small percentage of this DNA codes for proteins or functional RNA molecules. A chromosomal DNA molecule also contains three other types of functionally important nucleotide sequences: replication origins and telomeres allow the DNA molecule to be efficiently replicated, while a centromere attaches the daughter DNA molecules to the mitotic spindle, ensuring their accurate segregation to daughter cells during the M phase of the cell cycle.

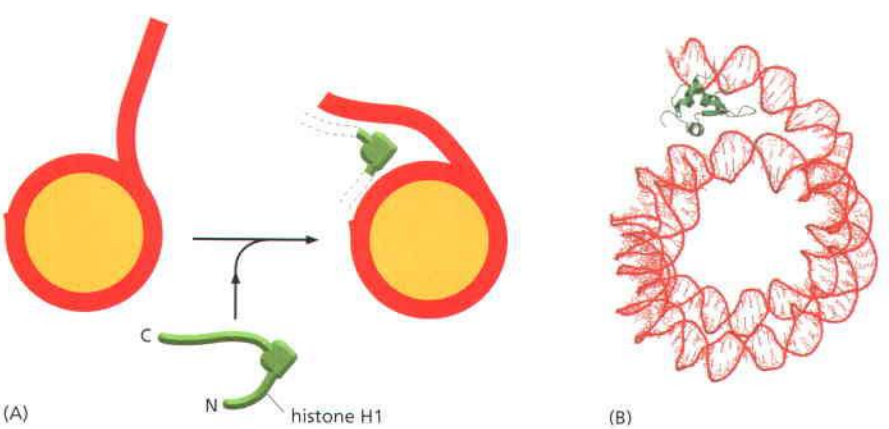


Figure 4–34 How the linker histone binds to the nucleosome. The position and structure of the globular region of histone H1 are shown. As indicated, this region constrains an additional 20 nucleotide pairs of DNA where it exits from the nucleosome core. This type of binding by H1 is thought to be important for forming the 30-nm chromatin fiber. The long C-terminal tail of histone H1 is also required for the high-affinity binding of H1 to chromatin, but neither its position or that of the N-terminal tail is known. (A) Schematic, (B) structure. (B, from D. Brown, T. Izard and T. Misteli, *Nat. Struct. Mol. Biol.* 13:250–255, 2006. With permission from Macmillan Publishers Ltd.)

The DNA in eucaryotes is tightly bound to an equal mass of histones, which form repeated arrays of DNA–protein particles called nucleosomes. The nucleosome is composed of an octameric core of histone proteins around which the DNA double helix is wrapped. Nucleosomes are spaced at intervals of about 200 nucleotide pairs, and they are usually packed together (with the aid of histone H1 molecules) into quasi-regular arrays to form a 30-nm chromatin fiber. Despite the high degree of compaction in chromatin, its structure must be highly dynamic to allow access to the DNA. There is some spontaneous DNA unwrapping and rewrapping in the nucleosome itself; however, the general strategy for reversibly changing local chromatin structure features ATP-driven chromatin remodeling complexes. Cells contain a large set of such complexes, which are targeted to specific regions of chromatin at appropriate times. The remodeling complexes collaborate with histone chaperones to allow nucleosome cores to be repositioned, reconstituted with different histones, or completely removed to expose the underlying DNA.

THE REGULATION OF CHROMATIN STRUCTURE

Having described how DNA is packaged into nucleosomes to create a chromatin fiber, we now turn to the mechanisms that create different chromatin structures in different regions of a cell's genome. We now know that mechanisms of this type are used to control many genes in eucaryotes. Most importantly, certain types of chromatin structure can be inherited; that is, the structure can be directly passed down from a cell to its decendents. Because the cell memory that results is based on an inherited protein structure rather than on a change in DNA sequence, this is a form of **epigenetic inheritance**. The prefix *epi* is Greek for "on"; this is appropriate, because epigenetics represents a form of inheritance that is superimposed on the genetic inheritance based on DNA (**Figure 4–35**).

In Chapter 7, we shall introduce the many different ways in which the expression of genes is regulated. There we discuss epigenetic inheritance in detail and present several distinct mechanisms that can produce it. Here, we are concerned with only one, that based on chromatin structure. We begin this section with an introduction to inherited chromatin structures and then describe the basis for them—the covalent modification of histones in nucleosomes. We shall see that these modifications serve as recognition sites for protein modules that bring specific protein complexes to the appropriate regions of chromatin, thereby producing specific effects on gene expression or inducing other biological functions. Through such mechanisms, chromatin structure plays a central role in the development, growth, and maintenance of eucaryotic organisms, including ourselves.

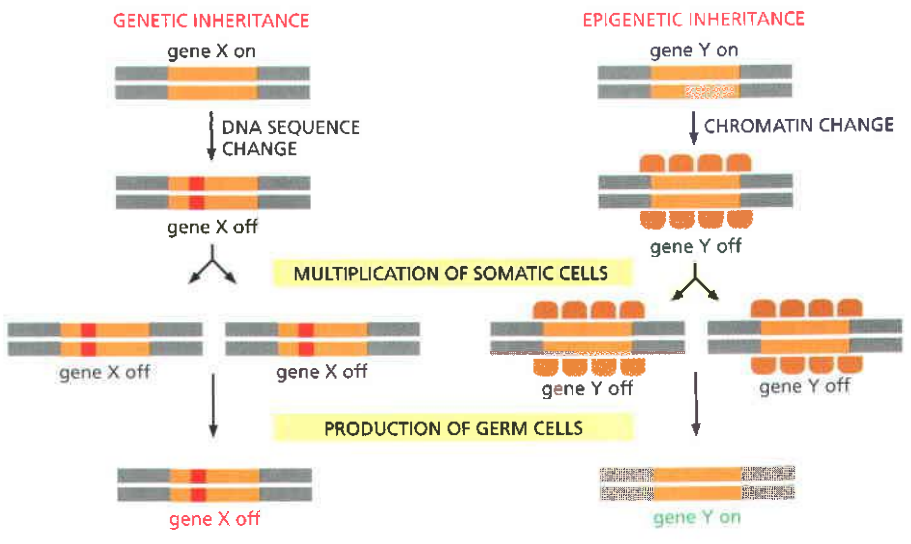


Figure 4–35 A comparison of genetic inheritance with an epigenetic inheritance based on chromatin structures. Genetic inheritance is based on the direct inheritance of DNA nucleotide sequences during DNA replication. DNA sequence changes are not only transmitted faithfully from a somatic cell to all of its descendents, but also through germ cells from one generation to the next. The field of genetics, reviewed in Chapter 8, is based on the inheritance of these changes between generations. The type of epigenetic inheritance shown here is based on other molecules bound to the DNA, and it is therefore less permanent than a change in DNA sequence; in particular, epigenetic information is usually (but not always) erased during the formation of eggs and sperm.

Only one epigenetic mechanism, that based on an inheritance of chromatin structures, is discussed in this chapter. Other epigenetic mechanisms are presented in Chapter 7, which focuses on the control of gene expression (see Figure 7–86).

Some Early Mysteries Concerning Chromatin Structure

Thirty years ago, histones were viewed as relatively uninteresting proteins. Nucleosomes were known to cover all of the DNA in chromosomes, and they were thought to exist to allow the enormous amounts of DNA in many eucaryotic cells to be packaged into compact chromosomes. Extrapolating from what was known in bacteria, many scientists believed that gene regulation in eucaryotes would simply bypass nucleosomes, treating them as uninvolved bystanders.

But there were reasons to challenge this view. Thus, for example, biochemists had determined that mammalian chromatin consists of an approximately equal mass of histone and non-histone proteins. This would mean that, *on average*, every 200 nucleotide pairs of DNA in our cells is associated with more than 1000 amino acids of non-histone proteins (that is, a mass of protein equivalent to the total mass of the histone octamer plus histone H1). We now know that many of these proteins bind to nucleosomes, and their abundance might suggest that histones are more than just packaging proteins.

A second reason to challenge the view that histones were inconsequential to gene regulation was based on the amazingly slow rate of evolutionary change in the sequences of the four core histones. The previously mentioned fact that there are only two amino acid differences in the sequence of mammalian and pea histone H4 implies that a change in almost any one of the 102 amino acids in H4 must be deleterious to these organisms. What type of process could make the life of an organism so sensitive to the exact structure of the nucleosome core that only two amino acids had changed in more than 500 million years of random variation followed by natural selection?

Last but not least, a combination of genetics and cytology had revealed that a particular form of chromatin silences the genes that it packages without regard to nucleotide sequence—and does so in a manner that is directly inherited by both daughter cells when a cell divides. It is to this subject that we turn next.

Heterochromatin Is Highly Organized and Unusually Resistant to Gene Expression

Light-microscope studies in the 1930s distinguished two types of chromatin in the interphase nuclei of many higher eucaryotic cells: a highly condensed form, called **heterochromatin**, and all the rest, which is less condensed, called **euchromatin**. Heterochromatin represents an especially compact form of chromatin (see Figure 4–9), and we are finally beginning to understand important aspects of its molecular properties. Although present in many locations along chromosomes, it is also highly concentrated in specific regions, most notably at the centromeres and telomeres introduced previously (see Figure 4–21). In a typical mammalian cell, more than ten percent of the genome is packaged in this way.

The DNA in heterochromatin contains very few genes, and those euchromatic genes that become packaged into heterochromatin are turned off by this type of packaging. However, we know now that the term *heterochromatin* encompasses several distinct types of chromatin structures whose common feature is an especially high degree of compaction. Thus, heterochromatin should not be thought of as encapsulating "dead" DNA, but rather as creating different types of compact chromatin with distinct features that make it highly resistant to gene expression for the vast majority of genes.

When a gene that is normally expressed in euchromatin is experimentally relocated into a region of heterochromatin, it ceases to be expressed, and the gene is said to be *silenced*. These differences in gene expression are examples of **position effects**, in which the activity of a gene depends on its position relative to a nearby region of heterochromatin on a chromosome. First recognized in *Drosophila*, position effects have now been observed in many eucaryotes, including yeasts, plants, and humans.

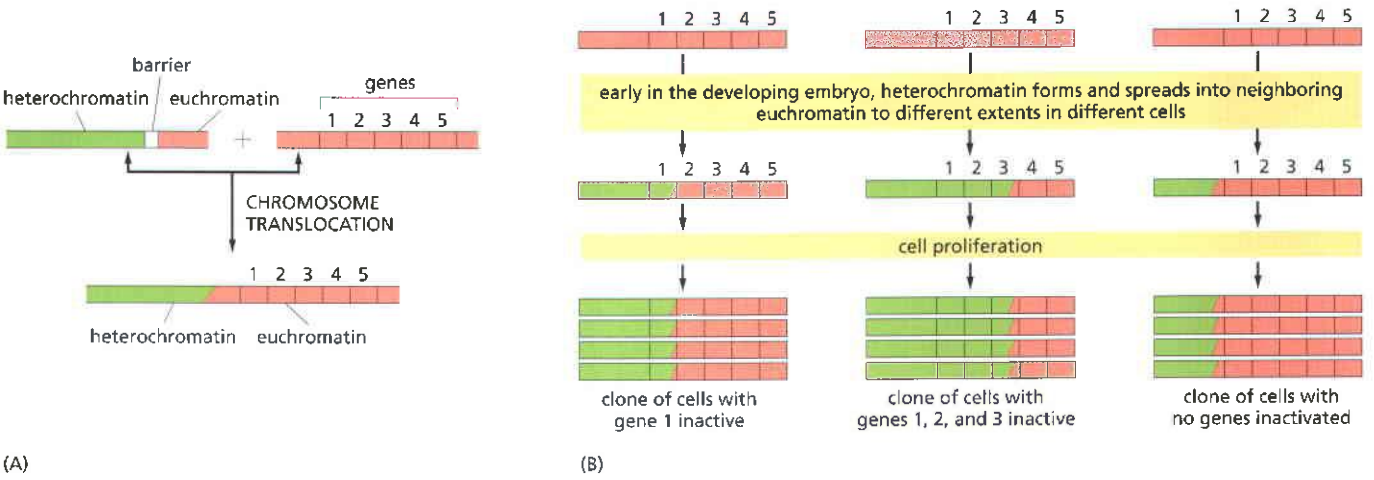


Figure 4–36 The cause of position effect variegation in *Drosophila*. (A) Heterochromatin *(green)* is normally prevented from spreading into adjacent regions of euchromatin *(red)* by special *barrier* DNA sequences, which we shall discuss shortly. In flies that inherit certain chromosomal rearrangements, however, this barrier is no longer present. (B) During the early development of such flies, heterochromatin can spread into neighboring chromosomal DNA, proceeding for different distances in different cells. This spreading soon stops, but the established pattern of heterochromatin is inherited, so that large clones of progeny cells are produced that have the same neighboring genes condensed into heterochromatin and thereby inactivated (hence the "variegated" appearance of some of these flies; see Figure 4–37). Although "spreading" is used to describe the formation of new heterochromatin close to previously existing heterochromatin, the term may not be wholly accurate. There is evidence that during expansion, heterochromatin can "skip over" some regions of chromatin, sparing the genes that lie within them from repressive effects.

The position effects associated with heterochromatin exhibit a feature called *position effect variegation*, which in retrospect provided critical clues concerning chromatin function. In *Drosophila*, chromosome breakage events that directly connect a region of heterochromatin to a region of euchromatin tend to inactivate the nearby euchromatic genes. The zone of inactivation spreads a different distance in different early cells in the fly embryo, but once the heterchromatic condition is established on a gene, it tends to be stably inherited by all of the cell's progeny (**Figure 4–36**). This remarkable phenomenon was first recognized through a detailed genetic analysis of the mottled loss of red pigment in the fly eye (**Figure 4–37**), but it shares many features with the extensive spread of heterochromatin that inactivates of one of the two X chromosomes in female mammals (see p. 473).

Extensive genetic screens have been carried out in *Drosophila*, as well as in fungi, in a search for gene products that either enhance or suppress the spread of heterochromatin and its stable inheritance—that is, for genes that when mutated serve as either enhancers or suppressors of position effect variegation. In this way, more than 50 genes have been identified that play a critical role in these processes. In recent years, the detailed characterization of the proteins produced by these genes has revealed that many are nonhistone chromosomal proteins that underlie a remarkable mechanism for eucaryotic gene control, one

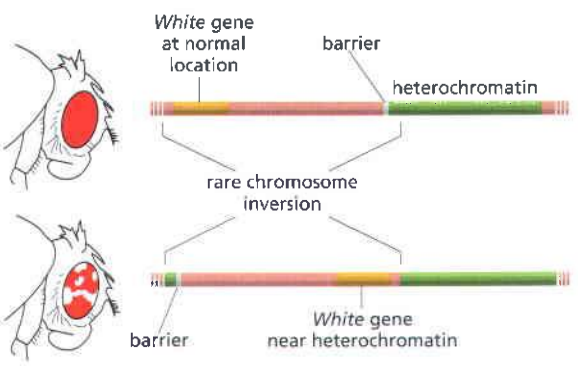
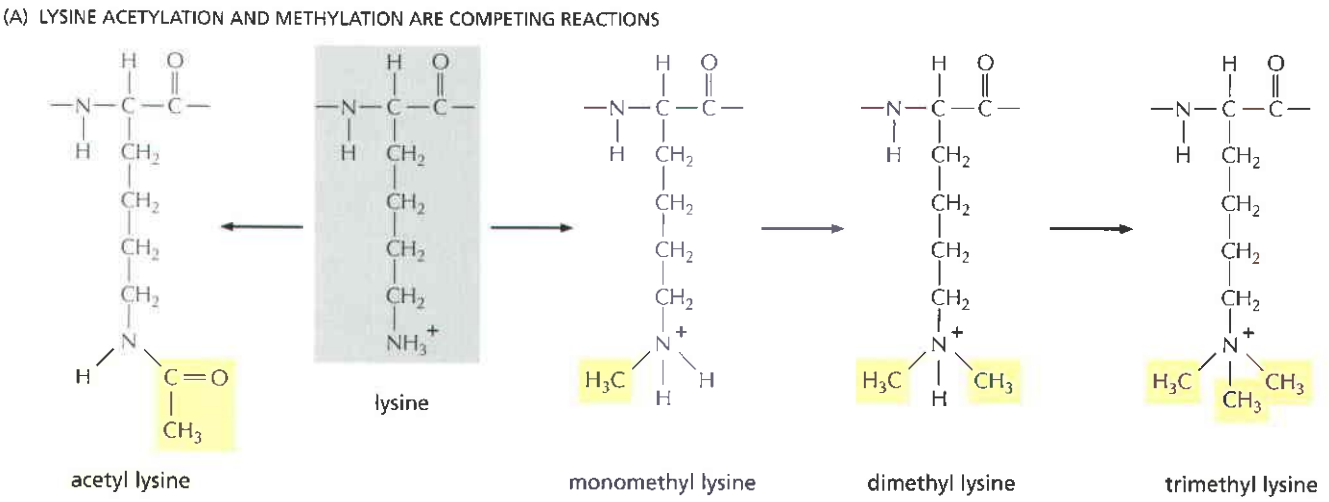


Figure 4–37 The discovery of position effects on gene expression. The *White* gene in the fruit fly *Drosophila* controls eye pigment production and is named after the mutation that first identified it. Wild-type flies with a normal *White* gene *(White$^+$)* have normal pigment production, which gives them *red* eyes, but if the *White* gene is mutated and inactivated, the mutant flies *(White$^-$)* make no pigment and have *white* eyes. In flies in which a normal *White$^+$* gene has been moved near a region of heterochromatin, the eyes are mottled, with both *red* and *white* patches. The *white* patches represent cell lineages in which the *White$^+$* gene has been silenced by the effects of the heterochromatin. In contrast, the *red* patches represent cell lineages in which the *White$^+$* gene is expressed. Early in development, when the heterochromatin is first formed, it spreads into neighboring euchromatin to different extents in different embryonic cells (see Figure 4–36). The presence of large patches of *red* and *white* cells reveals that the state of transcriptional activity, as determined by the packaging of this gene into chromatin in those ancestor cells, is inherited by all daughter cells.



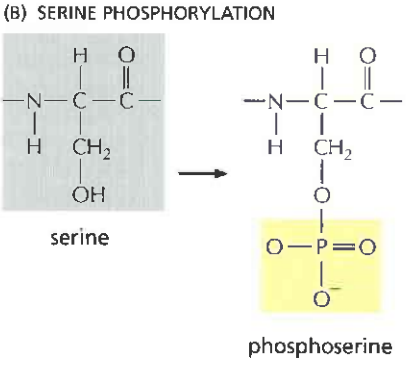


Figure 4–38 Some prominent types of covalent amino acid side-chain modifications found on nucleosomal histones. (A) Three different levels of lysine methylation are shown; each can be recognized by a different binding protein and thus each can have a different significance for the cell. Note that acetylation removes the plus charge on lysine, and that, most importantly, an acetylated lysine cannot be methylated, and *vice versa*. (B) Serine phosphorylation adds a negative charge to a histone. Modifications not shown here are the mono- or di-methylation of an arginine, the phosphorylation of a threonine, the addition of ADP-ribose to a glutamic acid, and the addition of a ubiquityl, sumoyl, or biotin group to a lysine.

that requires the precise amino acid sequences of the core histones. This mechanism of gene control therefore helps to explain the remarkably slow change in the histones over time.

The Core Histones Are Covalently Modified at Many Different Sites

The amino acid side chains of the four histones in the nucleosome core are subjected to a remarkable variety of covalent modifications, including the acetylation of lysines, the mono-, di-, and tri-methylation of lysines, and the phosphorylation of serines (**Figure 4–38**). A large number of these side-chain modifications occur on the eight relatively unstructured N-terminal "histone tails" that protrude from the nucleosome (**Figure 4–39**). However, there are also specific side-chain modifications on the nucleosome's globular core (**Figure 4–40**).

All of the above types of modifications are reversible. The modification of a particular amino acid side chain in a nucleosome is created by a specific enzyme, with most of these enzymes acting only on one or a few sites. A different enzyme is responsible for removing each side chain modification. Thus, for example, acetyl groups are added to specific lysines by a set of different histone acetyl transferases (HATs) and removed by a set of histone deacetylase complexes (HDACs). Likewise, methyl groups are added to lysine side chains by a set of different histone methyl transferases and removed by a set of histone demethylases. Each enzyme is recruited to specific sites on the chromatin at defined times in each cell's life history. For the most part, the initial recruitment of these enzymes depends on *gene regulatory proteins* that bind to specific DNA sequences along chromosomes, and these are produced at different times in the life of an organism, as described in Chapter 7. But in at least some cases, the covalent modifications on nucleosomes can persist long after the gene regulatory proteins that first induced them have disappeared, thereby carrying a memory in the cell of its developmental history. Very different patterns of covalent modifications are therefore found on different groups of nucleosomes, according to their exact position on a chromosome and the status of the cell.

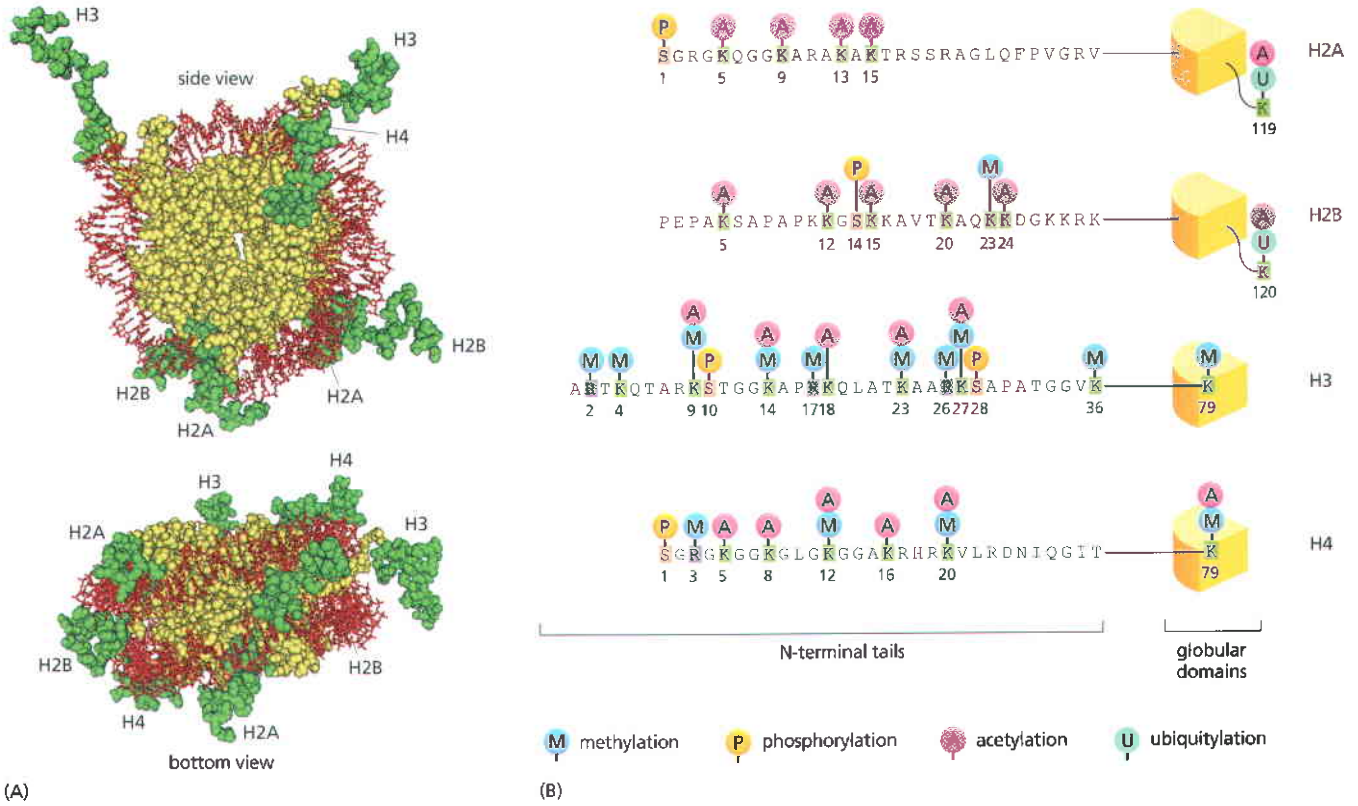


Figure 4–39 The covalent modification of core histone tails. (A) The structure of the nucleosome highlighting the location of the first 30 amino acids in each of its eight N-terminal histone tails *(green)*. (B) Well-documented modifications of the four histone core proteins are indicated. Although only a single symbol is used for methylation here (*M*), each lysine (*K*) or arginine (*R*) can be methylated in several different ways. Note also that some positions (e.g., lysine 9 of H3) can be modified either by methylation or by acetylation, but not both. Most of the modifications shown add a relatively small molecule onto the histone tails; the exception is ubiquitin, a 76 amino acid protein also used for other cell processes (see Figure 6–92). (Adapted from H. Santos-Rosa and C. Caldas, *Eur. J. Cancer* 41:2381–2402, 2005. With permission from Elsevier.)

The modifications of the histones are carefully controlled, and they have important consequences. The acetylation of lysines on the N-terminal tails tends to loosen chromatin structure, in part because adding an acetyl group to lysine removes its positive charge, thereby reducing the affinity of the tails for

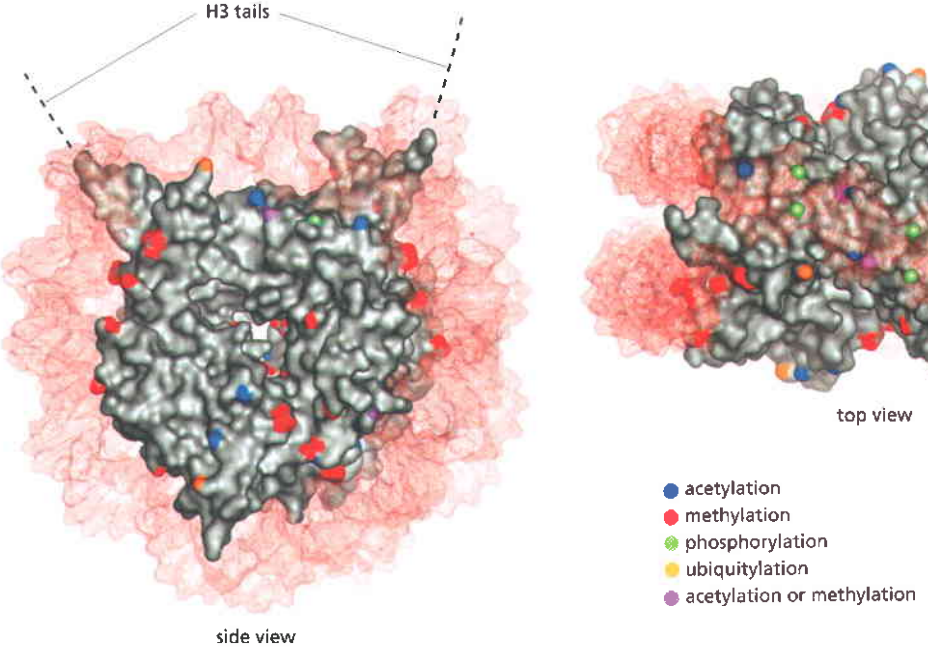


Figure 4–40 A map of histone modifications on the surface of the nucleosome core particle. As noted, the histone tails have been omitted here (compare with Figure 4–39). The functions of most of these core modifications are not yet known. (Adapted from M.S. Cosgrove, J.D. Boeke and C. Wolberger, *Nat. Struct. Mol. Biol.* 11:1037–1043, 2004. With permission from Macmillan Publishers Ltd.)

adjacent nucleosomes (see Figure 4–33). However, the most profound effect of the histone modifications is their ability to attract specific proteins to a stretch of chromatin that has been appropriately modified. These new proteins determine how and when genes will be expressed, as well as other biological functions. In this way, the precise structure of a domain of chromatin determines the expression of the genes packaged in it, and thereby the structure and function of the eucaryotic cell.

Chromatin Acquires Additional Variety Through the Site-Specific Insertion of a Small Set of Histone Variants

Despite the tight conservation of the amino acid sequences of the four core histones over hundreds of millions of years, eucaryotes also contain a few variant histones that assemble into nucleosomes. These histones are present in much smaller amounts than the major histones, and they have been less well conserved over long evolutionary times. Except for histone H4, variants exist for each of the core histones; some examples are shown in **Figure 4–41**.

The major histones are synthesized primarily during the S phase of the cell cycle (see Figure 17–4) and assembled into nucleosomes on the daughter DNA helices just behind the replication fork (see Figure 5–38). In contrast, most histone variants are synthesized throughout interphase. They are often inserted into already-formed chromatin, which requires a histone-exchange process catalyzed by the ATP-dependent chromatin remodeling complexes discussed previously. These remodeling complexes contain subunits that cause them to bind both to specific sites on chromatin and to histone chaperones that carry a particular variant. As a result, each histone variant is inserted into chromatin in a highly selective manner (see Figure 4–30).

The Covalent Modifications and the Histone Variants Act in Concert to Produce a "Histone Code" That Helps to Determine Biological Function

The number of possible distinct markings on an individual nucleosome is enormous. Even with the recognition that some of the covalent modifications are mutually exclusive (for example, it is not possible for a lysine to be both acetylated and methylated at the same time), and that other modifications are created together as a set, it is clear that thousands of combinations can exist. In addition, there is the further diversity created by nucleosomes that contain histone variants.

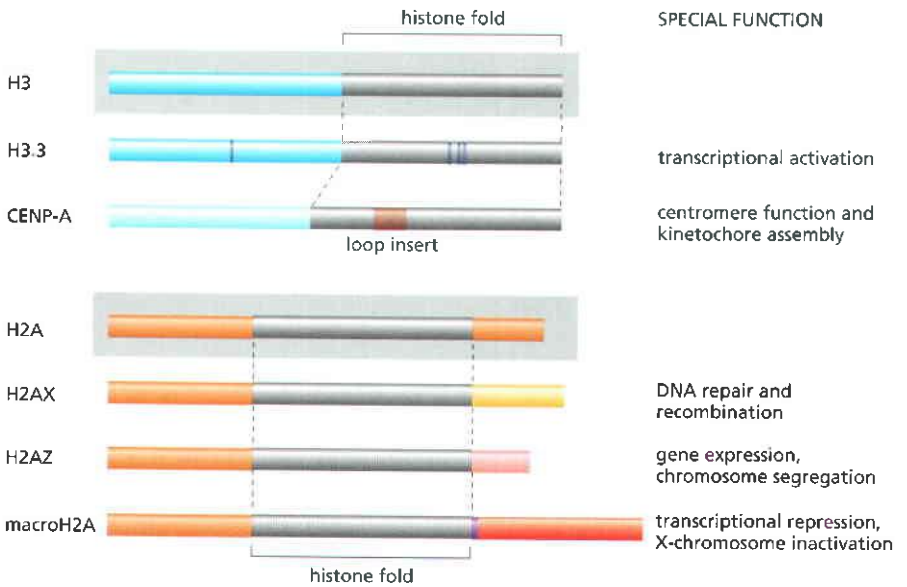


Figure 4–41 The structure of some histone variants compared with the major histone that they replace. These histones are inserted into nucleosomes at specific sites on chromosomes by ATP-dependent chromatin remodeling enzymes that act in concert with histone chaperones (see Figure 4–30). The CENP-A variant of histone H3 is discussed later in this chapter (see Figures 4–48 to 4–51); other variants are discussed in Chapter 7. The sequences that are colored differently in each variant are different from the corresponding sequence of the major histone. (Adapted from K. Sarma and D. Reinberg, *Nat. Rev. Mol. Cell. Biol.* 6:139–149, 2005. With permission from Macmillan Publishers Ltd.)

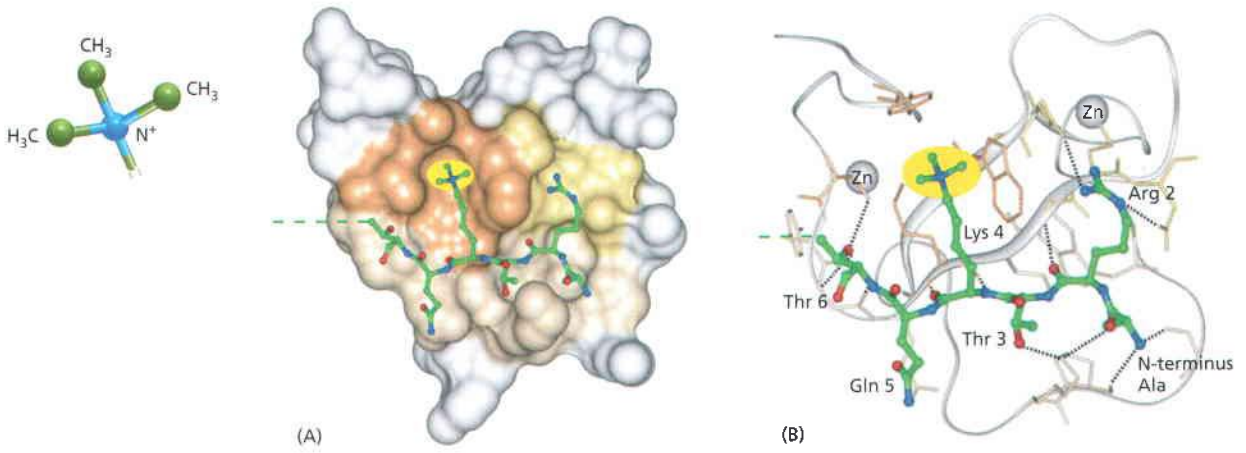


Figure 4–42 How each mark on a nucleosome is read. The structure of a protein module that specifically recognizes histone H3 trimethylated on lysine 4 is shown. (A) Space-filling model of an ING PHD domain bound to a histone tail (*green*, with the trimethyl group highlighted in *yellow*). (B) A ribbon model showing how the N-terminal six amino acids in the H3 tail are recognized. The *dashed lines* represent hydrogen bonds. This is one of many PHD domains that recognize methylated lysines on histones; different domains bind tightly to lysines located at different positions, and they can discriminate between a mono-, di-, and tri-methylated lysine. In a similar way, other small protein modules recognize specific histone side chains that have been marked with acetyl groups, phosphate groups, and so on. (Adapted from P.V. Pena et al., *Nature* 442:100–103, 2006. With permission from Macmillan Publishers Ltd.)

Many of the combinations appear to have a specific meaning for the cell because they determine how and when the DNA packaged in the nucleosomes is accessed, leading to the **histone code** hypothesis. For example, one type of marking signals that a stretch of chromatin has been newly replicated, another signals that the DNA in that chromatin has been damaged and needs repair, while many others signal when and how gene expression should take place. Small protein modules bind to specific marks, recognizing for example a tri-methylated lysine 4 on histone H3 (**Figure 4–42**). These modules are thought to act in concert with other modules as part of a *code-reader complex*, so as to allow particular combinations of markings on chromatin to attract additional protein complexes that execute an appropriate biological function at the right time (**Figure 4–43**).

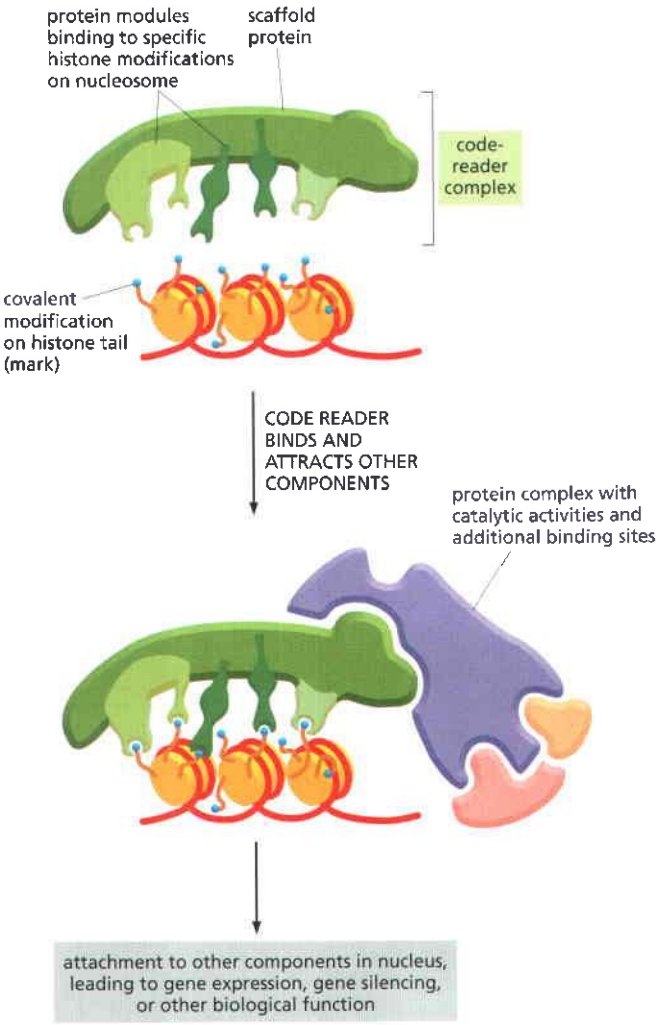


Figure 4–43 Schematic diagram showing how the histone code could be read by a code-reader complex. A large protein complex that contains a series of protein modules, each of which recognizes a specific histone mark, is schematically illustrated *(green)*. This "code-reader complex" will bind tightly only to a region of chromatin that contains several of the different histone marks that it recognizes. Therefore, only a specific combination of marks will cause the complex to bind to chromatin and attract additional protein complexes *(purple)* that catalyze a biological function.

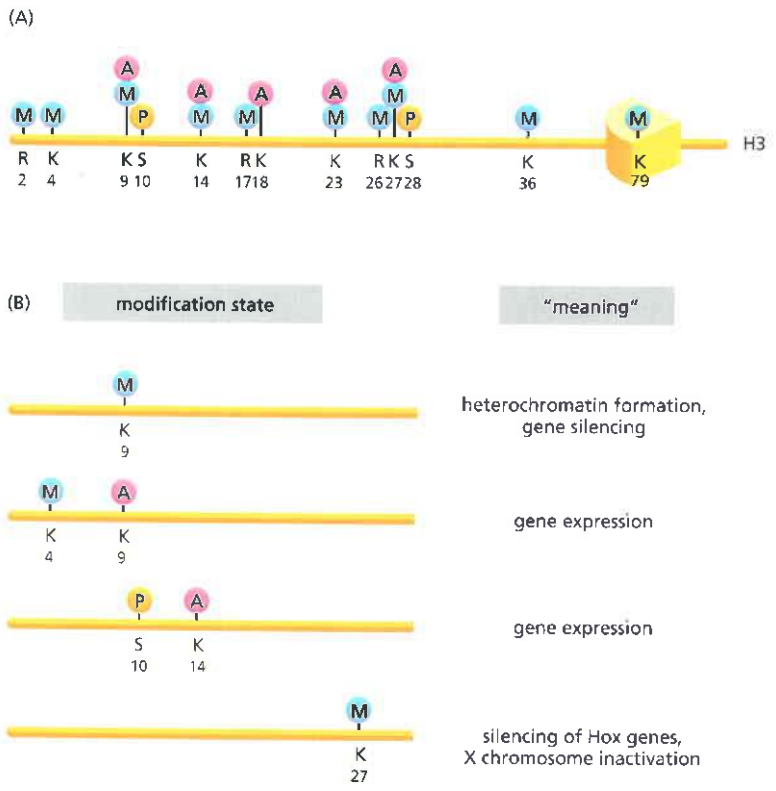


Figure 4–44 Some specific meanings of the histone code. (A) The modifications on the histone H3 N-terminal tail are shown, repeated from Figure 4–39. (B) The H3 tail can be marked by different combinations of modifications that convey a specific meaning to the stretch of chromatin where this combination occurs. Only a few of the meanings are known, including the four examples shown. To focus on just one example, the trimethylation of lysine 9 attracts the heterochromatin-specific protein HP1, which induces a spreading wave of further lysine 9 trimethylation followed by further HP1 binding, according to the general scheme that will be illustrated shortly (see Figure 4–46). Not shown is the fact that, as just implied (see Figure 4–43), reading the histone code generally involves the joint recognition of marks at other sites on the nucleosome along with the indicated H3 tail recognition. In addition, specific levels of methylation (mono-, di-, or tri-methyl groups) are required, as in Figure 4–42.

The marks on nucleosomes due to covalent additions to histones are dynamic, being constantly removed and added at rates that depend on their chromosomal locations. Because the histone tails extend outward from the nucleosome core and are likely to be accessible even when chromatin is condensed, they would seem to provide an especially suitable format for creating marks in a form that can be readily altered as a cell's needs change. Although much remains to be learned about the meaning of the many different histone code combinations, a few well-studied examples of the information that can be encoded in the histone H3 tail are listed in Figure 4–44.

A Complex of Code-reader and Code-writer Proteins Can Spread Specific Chromatin Modifications for Long Distances Along a Chromosome

The phenomenon of position effect variegation described previously requires that at least some modified forms of chromatin have the ability to spread for substantial distances along a chromosomal DNA molecule (see Figure 4–36). How is this possible?

The enzymes that modify (or remove modifications from) the histones in nucleosomes are part of multisubunit complexes. They can initially be brought to a particular region of chromatin by one of the sequence-specific DNA-binding proteins (gene regulatory proteins) discussed in Chapters 6 and 7 (for a specific example, see Figure 7–87). But after a modifying enzyme "writes" its mark on one or a few neighboring nucleosomes, events that resemble a chain reaction can ensue. In this case, the "code-writer" enzyme works in concert with a code-reader protein located in the same protein complex. This second protein contains a code-reader module that recognizes the mark and binds tightly to the newly modified nucleosome (see Figure 4–42), positioning its attached writer enzyme near an adjacent nucleosome. Through many such read–write cycles, the reader protein can carry the writer enzyme along the DNA—spreading the mark in a hand-over-hand manner along the chromosome (**Figure 4–45**).

In reality, the process is more complicated than the scheme just described. Both readers and writers are part of a protein complex that is likely to contain

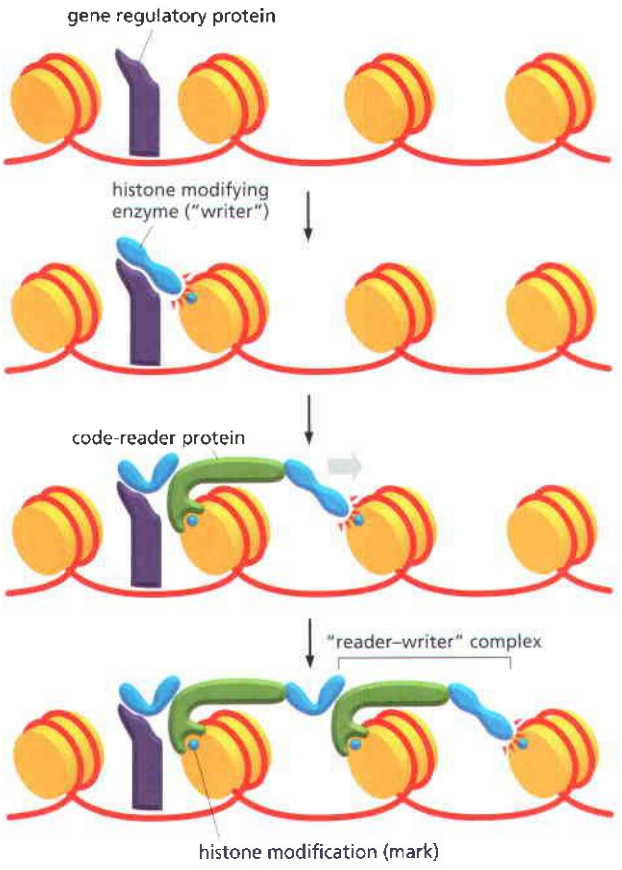


Figure 4–45 How the recruitment of a code-reader–writer complex can spread chromatin changes along a chromosome. The code-writer is an enzyme that creates a specific modification on one or more of the four nucleosomal histones. After its recruitment to a specific site on a chromosome by a gene regulatory protein, the writer collaborates with a code-reader protein to spread its mark from nucleosome to nucleosome by means of the indicated reader–writer complex. For this mechanism to work, the reader must recognize the same histone modification mark that the writer produces (see also Figure 4–43).

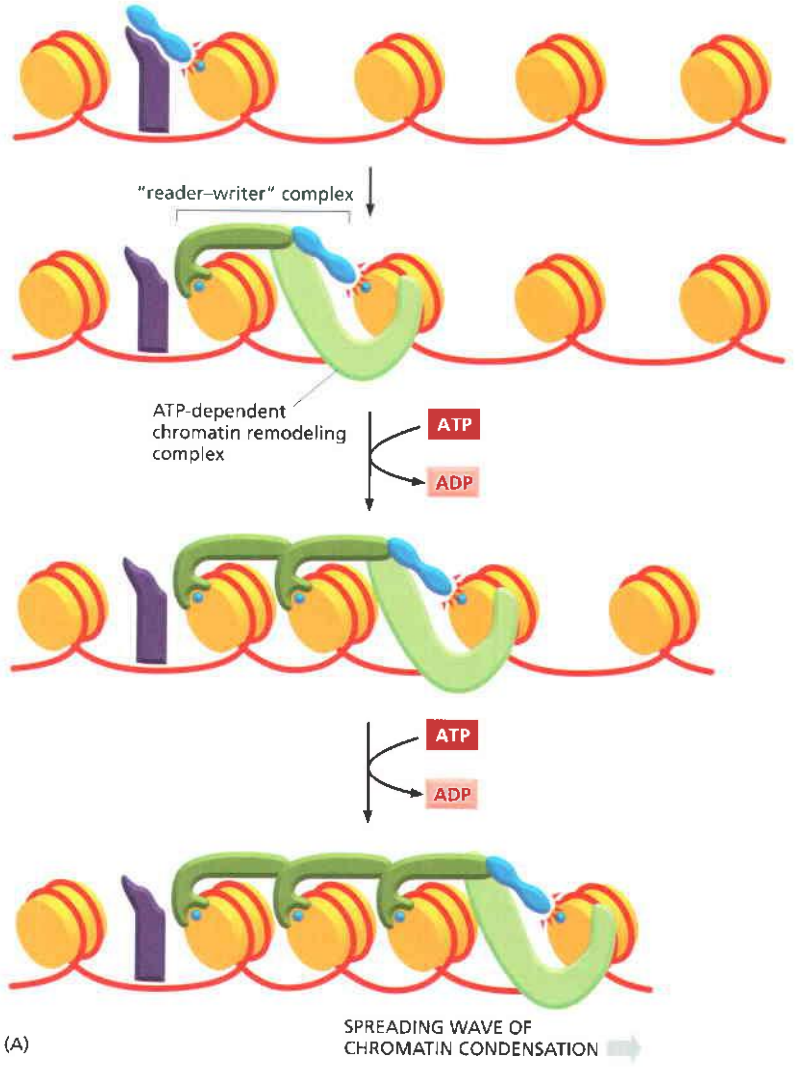
multiple readers and writers, and to require multiple marks on the nucleosome to spread. Moreover, many of these reader–writer complexes also contain an ATP-dependent chromatin remodeling protein, and the reader, writer, and remodeling proteins work in concert to either decondense or condense long stretches of chromatin as the reader moves progressively along the nucleosome-packaged DNA (**Figure 4–46**).

Some idea of the complexity of the processes just described can be derived from the results of genetic screens for mutant genes that either enhance or suppress the spreading and stability of heterochromatin in tests for position effect variegation in *Drosophila* (see Figure 4–37). As pointed out previously, more than 50 such genes are known, and most of them are likely to function as subunits in one or more reader–writer-remodeling protein complexes.

Barrier DNA Sequences Block the Spread of Reader–Writer Complexes and thereby Separate Neighboring Chromatin Domains

The above mechanism for spreading chromatin structures raises a potential problem. Inasmuch as each chromosome consists of one continuous, very long DNA molecule, what prevents a cacophony of confusing cross-talk between adjacent chromatin domains of different structure and function? Early studies of position effect variegation had suggested an answer: the existence of specific DNA sequences that separate one chromatin domain from another (see Figure 4–37). Several such *barrier* sequences have now been identified and characterized through the use of genetic engineering techniques that allow specific regions of DNA sequence to be deleted or added to chromosomes.

For example, a sequence called HS4 normally separates the active chromatin domain that contains the β-globin locus from an adjacent region of silenced, condensed chromatin in erythrocytes (see Figure 7–61). If this sequence is deleted, the β-globin locus is invaded by condensed chromatin. This chromatin silences the genes it covers, and it spreads to a different extent in different cells, causing a pattern of position effect variegation similar to that



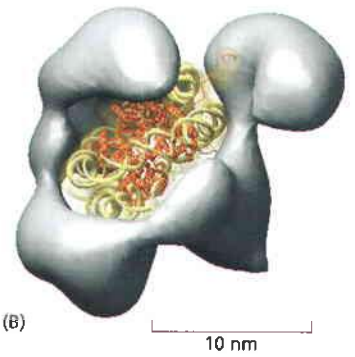


Figure 4–46 How a complex containing reader–writer and ATP-dependent chromatin remodeling proteins can spread chromatin changes along a chromosome. (A) A spreading wave of chromatin condensation. This mechanism is identical to that in Figure 4–45, except that the reader–writer complex collaborates with an ATP-dependent chromatin remodeling protein (see Figure 4–29) to reposition nucleosomes and pack them into highly condensed arrays. This is a highly simplified view of the mechanism known to be able to spread a major form of heterochromatin for long distances along chromosomes (see Figure 4–36). The heterochromatin-specific protein HP1 plays a major role in that process. HP1 binds to trimethyl lysine 9 on histone H3, and it remains associated with the condensed chromatin as one of the readers in a reader–writer-remodeling complex that, while incompletely understood, is considerably more intricate than that shown here. (B) The actual structure of a chromatin reader-remodeling complex, showing how it is thought to interact with a nucleosome. Modeled in *gray* is the yeast RSC complex, which contains 15 subunits—including an ATP-dependent chromatin remodeling protein and at least 4 subunits with code-reader domains. (B, from A.E. Leschziner et al., *Proc. Natl Acad. Sci. U.S.A.* 104:4913–4918, 2007. With permission from National Academy of Sciences.)

observed in *Drosophila*. As described in Chapter 7, this invasion has dire consequences: the globin genes are poorly expressed, and individuals who carry such a deletion have a severe form of anemia.

The HS4 sequence is often added to both ends of a gene that is experimentally inserted into a mammalian genome, in order to protect that gene from the silencing caused by spreading heterochromatin. Analysis of this barrier sequence reveals that it contains a cluster of binding sites for histone acetylase enzymes. Since the acetylation of a lysine side chain is incompatible with the methylation of the same side chain, histone acetylases and histone deacetylases are logical candidates for the formation of barriers on the DNA that block the spread of different forms of chromatin (**Figure 4–47**). However, several other types of chromatin modifications are known that can also protect genes from silencing.

The Chromatin in Centromeres Reveals How Histone Variants Can Create Special Structures

The presence of nucleosomes carrying histone variants is thought to produce marks in chromatin that are unusually long lasting. Consider, for example, the formation and inheritance of the chromatin that forms on centromeres, the DNA region of each chromosome required for the orderly segregation of the chromosomes into daughter cells each time a cell divides (see Figure 4–21). In many complex organisms, including humans, each centromere is embedded in a stretch of special *centric heterochromatin* that persists throughout interphase, even though the centromere-mediated movement of DNA occurs only during

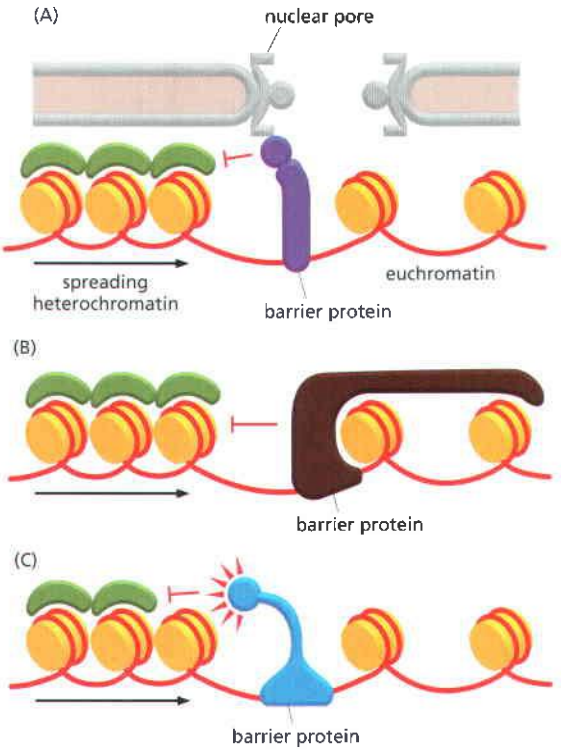


Figure 4–47 Some mechanisms of barrier action. These models are derived from different analyses of barrier action, and a combination of several of them may function at any one site. (A) The tethering of a region of chromatin to a large fixed site, such as the nuclear pore complex illustrated here, can form a barrier that stops the spread of heterochromatin. (B) The tight binding of barrier proteins to a group of nucleosomes can compete with heterochromatin spreading. (C) By recruiting a group of highly active histone-modifying enzymes, barriers can erase the histone marks that are required for heterochromatin to spread. For example, a potent acetylation of lysine 9 on histone H3 will compete with lysine 9 methylation, thereby preventing the HP1 protein binding needed to form some forms of heterochromatin (see Figure 4–46). (Based on A.G. West and P. Fraser, *Hum. Mol. Genet.* 14:R101–R111, 2005. With permission from Oxford University Press.)

mitosis. This chromatin contains a centromere-specific variant H3 histone, known as CENP-A (see Figure 4–41), plus additional proteins that pack the nucleosomes into particularly dense arrangements and form the kinetechore, the special structure required for attachment of the mitotic spindle.

A specific DNA sequence of approximately 125 nucleotide pairs is sufficient to serve as a centromere in the yeast *S. cerevisiae*. Despite its small size, more than a dozen different proteins assemble on this DNA sequence; the proteins include the CENP-A histone H3 variant, which, along with the three other core histones, forms a centromere-specific nucleosome. The additional proteins at the yeast centromere attach this nucleosome to a single microtubule from the yeast mitotic spindle (**Figure 4–48**).

The centromeres in more complex organisms are considerably larger than those in budding yeasts. For example, fly and human centromeres extend over hundreds of thousands of nucleotide pairs and do not seem to contain a centromere-specific DNA sequence. These centromeres largely consist of short, repeated DNA sequences, known as *alpha satellite DNA* in humans. But the same repeat sequences are also found at other (non-centromeric) positions on

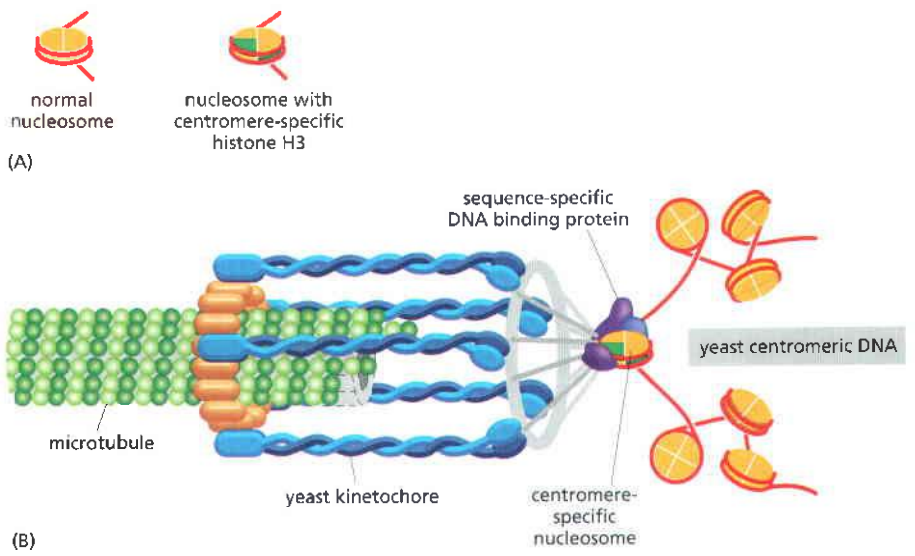


Figure 4–48 A model for the structure of a simple centromere. In the yeast *Saccharomyces cerevisiae*, a special centromeric DNA sequence assembles a single nucleosome in which two copies of an H3 variant histone (called CENP-A in most organisms) replaces the normal H3. Peptide sequences unique to this variant histone (see Figure 4–41) then help to assemble additional proteins, some of which form a kinetochore. This kinetochore is unusual in capturing only a single microtubule; humans have much larger centromeres and form kinetochores that can capture 20 or more microtubules (see Figure 4–50). The kinetochore is discussed in detail in Chapter 17. (Adapted from A. Joglekar et al., *Nat. Cell Biol.* 8:381–383, 2006. With permission from Macmillan Publishers Ltd.)

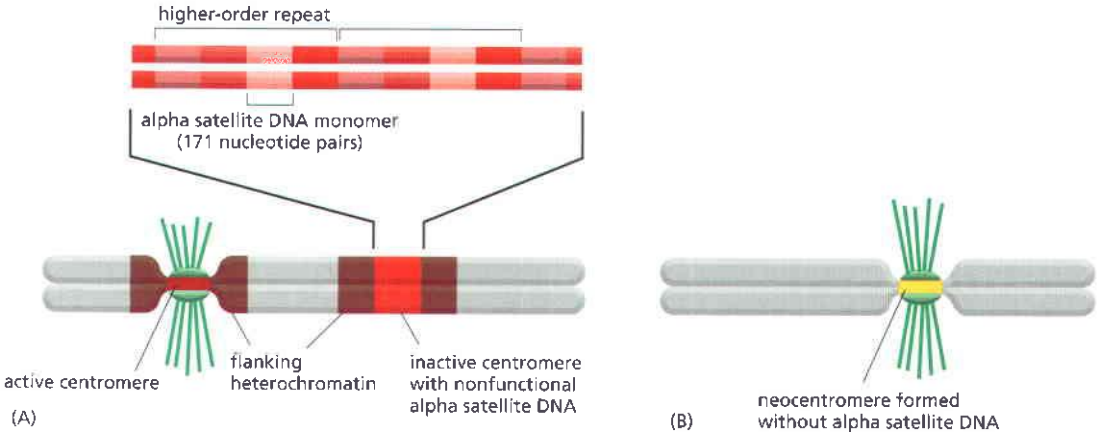


Figure 4–49 Evidence for the plasticity of human centromere formation. (A) A series of A-T-rich alpha satellite DNA sequences are repeated many thousands of times at each human centromere *(red)*, surrounded by pericentric heterochromatin *(brown)*. However, due to an ancient chromosome breakage and rejoining event, some human chromosomes contain two blocks of alpha satellite DNA, each of which presumably functioned as a centromere in its original chromosome. Usually, these dicentric chromosomes are not stably propagated because they attach improperly to the spindle and are broken apart during mitosis. In chromosomes that do survive, however, one of the centromeres has somehow inactivated, even though it contains all the necessary DNA sequences. This allows the chromosome to be stably propagated. (B) In a small fraction (1/2000) of human births, extra chromosomes are observed in cells of the offspring. Some of these extra chromosomes, which have formed from a breakage event, lack alpha satellite DNA altogether, yet new centromeres (neocentromeres) have arisen from what was originally euchromatic DNA.

chromosomes, indicating that they are not sufficient to direct centromere formation. Most strikingly, in some unusual cases, new human centromeres (called neocentromeres) have been observed to form spontaneously on fragmented chromosomes. Some of these new positions were originally euchromatic and lack alpha satellite DNA altogether (**Figure 4–49**).

It therefore seems that centromeres in complex organisms are defined by an assembly of proteins, instead of by a specific DNA sequence. When antibodies that stain specific modified nucleosomes are used to examine the stretched chromosome fibers from centromeres, one observes striking alternation of two modified forms of chromatin (**Figure 4–50**). It appears that this arrangement allows the centric heterochromatin to fold so as to position the CENP-A-containing nucleosomes on the outside of the mitotic chromosome, where they bind the set of proteins that form the kinetechore plates. These plates in turn capture a group of microtubules from the mitotic spindle in order to partition the chromosomes accurately, as described in Chapter 17.

Chromatin Structures Can Be Directly Inherited

To explain the above observations, it has been proposed that *de novo* centromere formation requires an initial seeding event, involving the formation of a specialized DNA–protein structure that contains nucleosomes formed with the CENP-A variant of histone H3. In humans, this seeding event happens more readily on arrays of alpha satellite DNA than on other DNA sequences. The H3–H4 tetramers from each nucleosome on the parental DNA helix are directly inherited by the daughter DNA helices at a replication fork (see Figure 5–38). Therefore, once a set of CENP-A-containing nucleosomes has been assembled on a stretch of DNA, it is easy to understand how a new centromere could be generated in the same place on both daughter chromosomes following each round of cell division (**Figure 4–51**).

The plasticity of centromeres may provide an important evolutionary advantage. We have seen that chromosomes evolve in part by breakage and rejoining events (see Figure 4–18). Many of these events produce chromosomes with two centromeres, or chromosome fragments with no centromeres at all. Although rare, both the inactivation of centromeres and their ability to be activated *de novo*

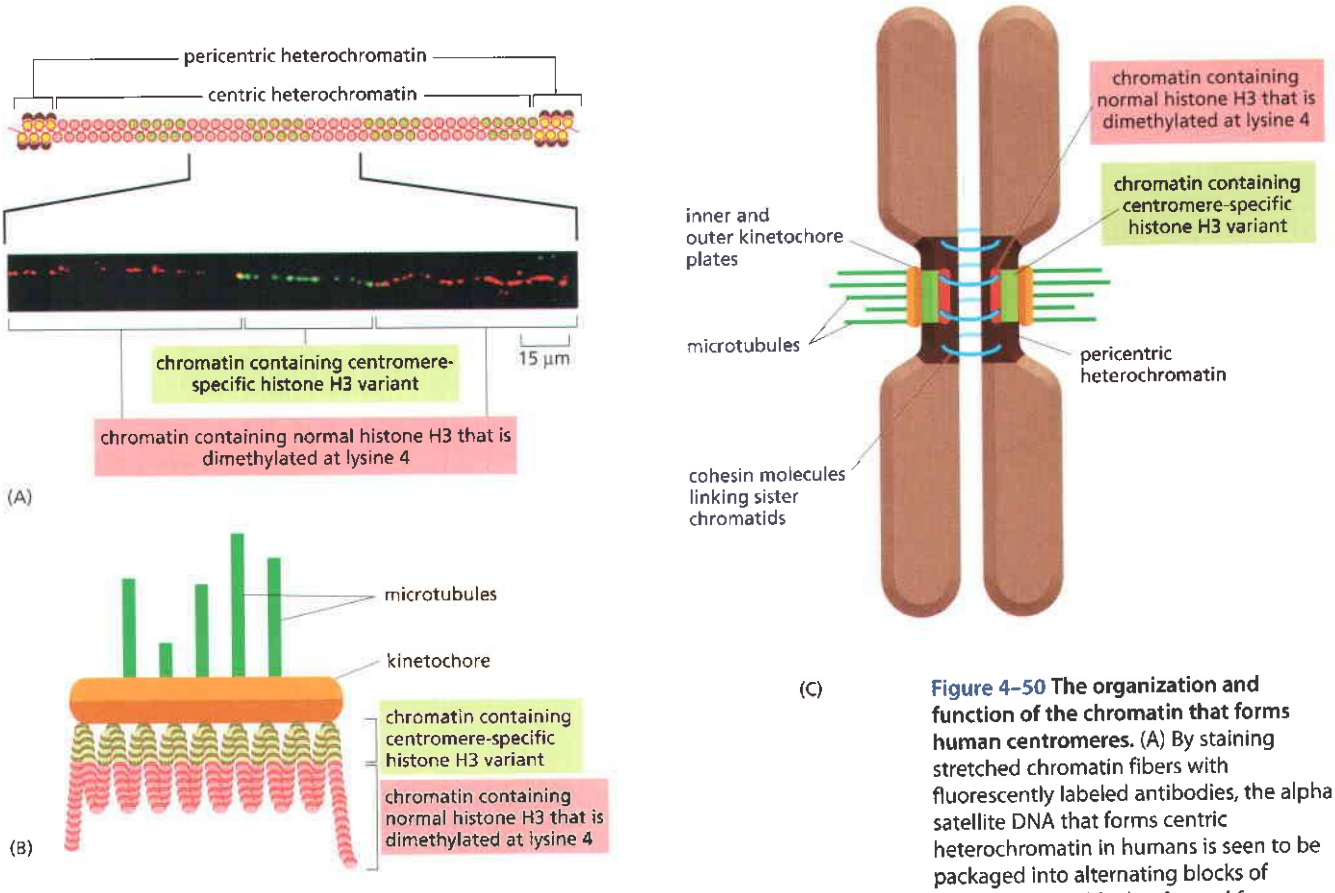


Figure 4–50 The organization and function of the chromatin that forms human centromeres. (A) By staining stretched chromatin fibers with fluorescently labeled antibodies, the alpha satellite DNA that forms centric heterochromatin in humans is seen to be packaged into alternating blocks of chromatin. One block is formed from a long string of nucleosomes containing the CENP-A H3 variant histone *(green)*; the other block contains nucleosomes that are specially marked with a dimethyl lysine 4 *(red)*. Each block is more than a thousand nucleosomes long. (B) A model for the organization of the two types of centric heterochromatin. As in yeast, the nucleosomes that contain the H3 variant histone form the kinetochore. (C) The arrangement of the centric and pericentric heterochromatin on a human metaphase chromosome, as determined by fluorescence microscopy using the same antibodies as in (A). (Adapted from B.A. Sullivan and G.H. Karpen, *Nat. Struct. Mol. Biol.* 11:1076–1083, 2004. With permission from Macmillan Publishers Ltd.)

may occasionally allow newly formed chromosomes to be maintained stably, thereby facilitating the process of chromosome evolution.

There are some striking similarities between the formation and maintenance of centromeres and the formation and maintenance of other regions of heterochromatin. In particular, the entire centromere forms as an all-or-none entity, suggesting a highly cooperative addition of proteins after a seeding event. Moreover, once formed, the structure seems to be directly inherited on the DNA as part of each round of chromosome replication.

Chromatin Structures Add Unique Features to Eucaryotic Chromosome Function

Although a great deal remains to be learned about the functions of different chromatin structures, the packaging of DNA into nucleosomes was probably crucial for the evolution of eucaryotes like ourselves. Complex multicellular organisms would appear to be possible only if the cells in different lineages can specialize by changing the accessibility and responsiveness of many hundreds of genes to genetic readout. As described in Chapter 22, each cell has a stored memory of its past developmental history in the regulatory circuits that control its many genes.

Although bacteria also require cell memory mechanisms, the complexity of the memory circuits required by higher eucaryotes is unprecedented. The packaging of selected regions of eucaryotic genomes into different forms of chromatin makes possible a type of cell memory mechanism that is not available to bacteria. The crucial feature of this uniquely eucaryotic form of gene regulation is the storage of the memory of the state of a gene on a gene-by-gene basis—in the form of local chromatin structures that can persist for various lengths of time. At one extreme are structures like centric heterochromatin that, once established, are stably inherited from one cell generation to the next (see Figure

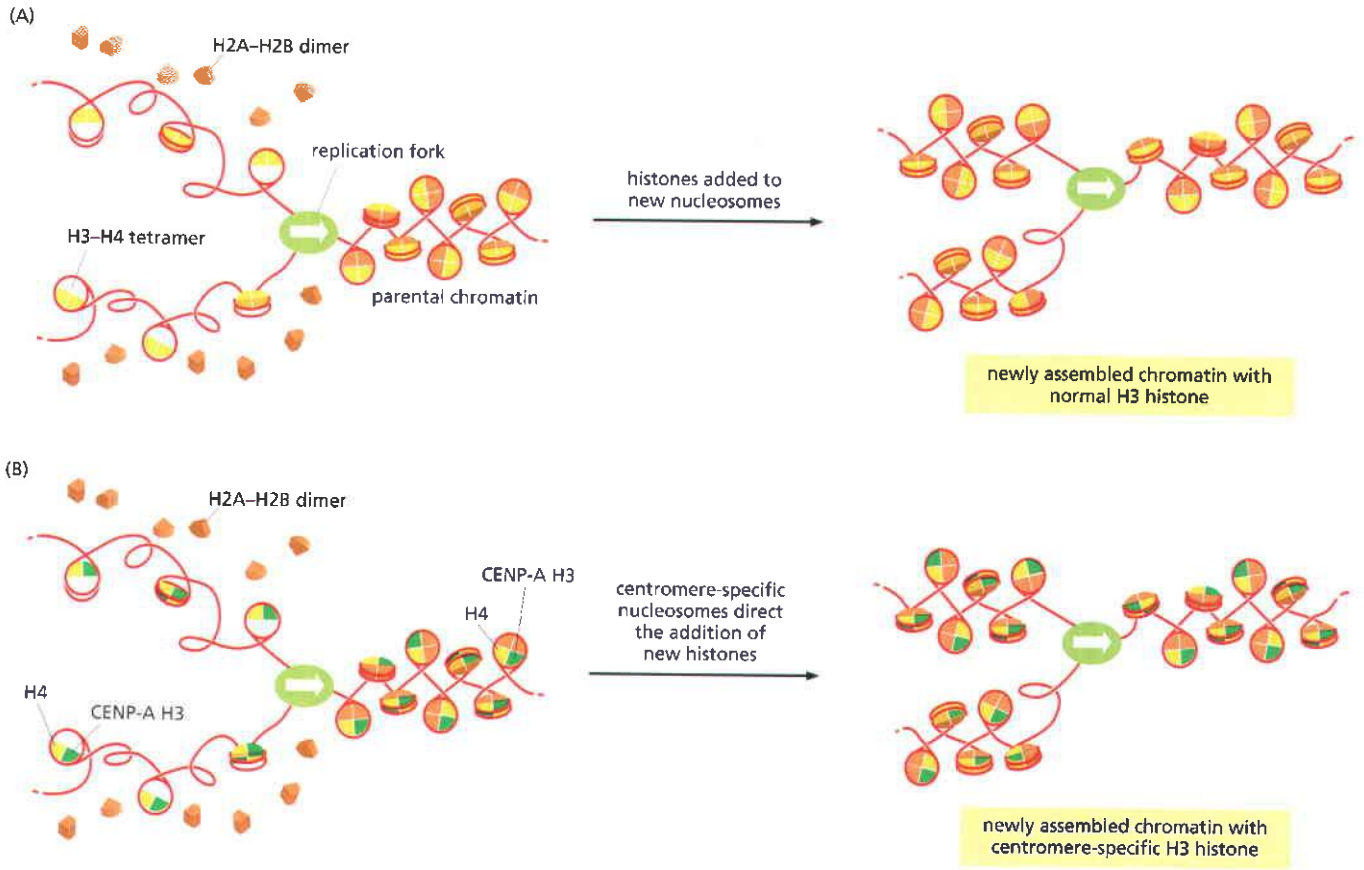


Figure 4–51 A model for the direct inheritance of centromeric heterochromatin. (A) The normal assembly of chromatin on the two daughter DNA helices produced at a replication fork requires the deposition of H2A–H2B dimers onto directly inherited H3–H4 tetramers, as well as the assembly of new histone octamers (see Figure 5–38 for details). (B) At a centromere, the inheritance of H3 variant–H4 tetramers seeds the formation of new histone octamers that likewise contain the variant H3 histone. A similar seeding process could cause the adjacent blocks of centric heterochromatin (containing H3 modified at dimethyl lysine 4; see Figure 4–50) to be inherited. Although the details are not known, the seeding process is likely to involve other centromeric proteins that are inherited along with the nucleosomes (see Figure 4–52).

4–51). Closely related mechanisms that are likewise based on the direct inheritance of parental forms of chromatin by the daughter DNA helices behind the replication fork are thought to be responsible for other types of condensed chromatin (**Figure 4–52**). For example, the permanently silenced, classical type

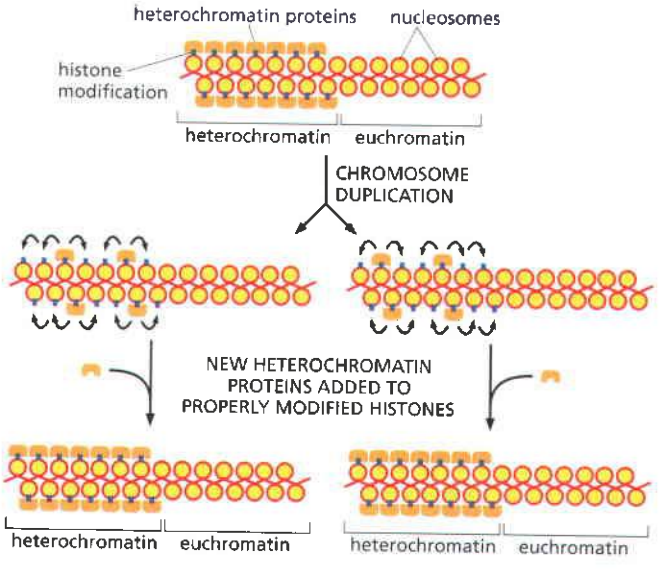


Figure 4–52 How the packaging of DNA in chromatin can be inherited during chromosome replication. In this model, some of the specialized chromatin components are distributed to each daughter chromosome after DNA duplication, along with the specially marked nucleosomes that they bind. After DNA replication, the inherited nucleosomes that are specially modified, acting in concert with the inherited chromatin components, change the pattern of histone modification on the newly formed daughter nucleosomes nearby. This creates new binding sites for the same chromatin components, which then assemble to complete the structure. The latter process is likely to involve code reader–writer-remodeling complexes operating in a manner similar to that previously illustrated in Figure 4–46.

of heterochromatin contains the HP1 protein, whereas the condensed chromatin that coats important developmental regulatory genes is maintained by the polycomb group of proteins. The latter type of heterochromatin silences a large number of genes that encode gene regulatory proteins early in embryonic development, covering a total of about 2 percent of the human genome, and it is removed only when each individual gene is needed by the developing organism (discussed in Chapter 22). Although other types of inherited chromatin structures exist, it is not yet clear how many different types there are: the number could certainly exceed 10 (see p. 238). The fundamental importance of this mechanism for distinguishing different genes is schematically represented in (**Figure 4–53**).

Other forms of chromatin can have a shorter lifetime, much less than the division time of the cell; however, many have a built-in persistence that helps to mediate biological function.

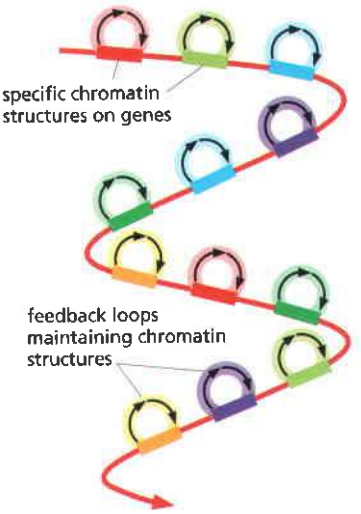


Figure 4–53 Schematic illustration of cell memory stored as chromatin-based epigenetic information in the genes of eucaryotes. Genes in eucaryotic cells can be packaged into a large variety of different chromatin structures, indicated here by different colors. At least some of these chromatin structures have a special effect on gene expression that can be directly inherited as epigenetic information when a cell divides. This allows some of the gene regulatory proteins that create different gene states to act only once, inasmuch as the state can be remembered after the regulatory protein is gone. Epigenetic information can also be stored in networks of signaling molecules that control gene expression (see Figure 7–86).

Summary

Despite the uniform assembly of chromosomal DNA into nucleosomes, a large variety of different chromatin structures are possible in eucaryotic organisms. This variety is based on a large set of reversible covalent modifications of the four histones in the nucleosome core. These modifications include the mono-, di-, and tri-methylation of many different lysine side chains, an important reaction that is incompatible with the acetylation of the same lysines. Specific combinations of the modifications mark each nucleosome with a histone code. The histone code is read when protein modules that are part of a larger protein complex bind to the modified nucleosomes in a region of chromatin. These code-reader proteins then attract additional proteins that catalyze biologically relevant functions.

Some code-reader protein complexes contain a histone-modifying enzyme, such as a histone methylase, that "writes" the same mark that the code-reader recognizes. A reader–writer-remodeling complex of this type can spread a specific form of chromatin for long distances along a chromosome. In particular, large regions of condensed heterochromatin are thought to be formed in this way. Heterochromatin is commonly found around centromeres and near telomeres, but it is also present at many other positions in chromosomes. The tight packaging of DNA into heterochromatin usually silences the genes within it.

The phenomenon of position effect variegation provides good evidence for the direct inheritance of condensed forms of chromatin by the daughter DNA helices formed at a replication fork, and a similar mechanism appears to be responsible for maintaining the specialized chromatin at centromeres. More generally, the ability to transmit specific chromatin structures from one cell generation to the next provides the basis for an epigenetic cell memory process that is likely to be critical for maintaining the complex set of different cell states required by complex multicellular organisms.

THE GLOBAL STRUCTURE OF CHROMOSOMES

Having discussed the DNA and protein molecules from which the 30-nm chromatin fiber is made, we now turn to the organization of the chromosome on a more global scale. As a 30-nm fiber, the typical human chromosome would still be 0.1 cm in length and able to span the nucleus more than 100 times. Clearly, there must be a still higher level of folding, even in interphase chromosomes. Although its molecular basis is still largely a mystery, this higher-order packaging almost certainly involves the folding of the 30-nm fiber into a series of loops and coils. This chromatin packing is fluid, frequently changing in response to the needs of the cell.

We shall begin by describing some unusual interphase chromosomes that can be easily visualized, inasmuch as certain features of these exceptional cases are thought to be representative of all interphase chromosomes. Moreover, they

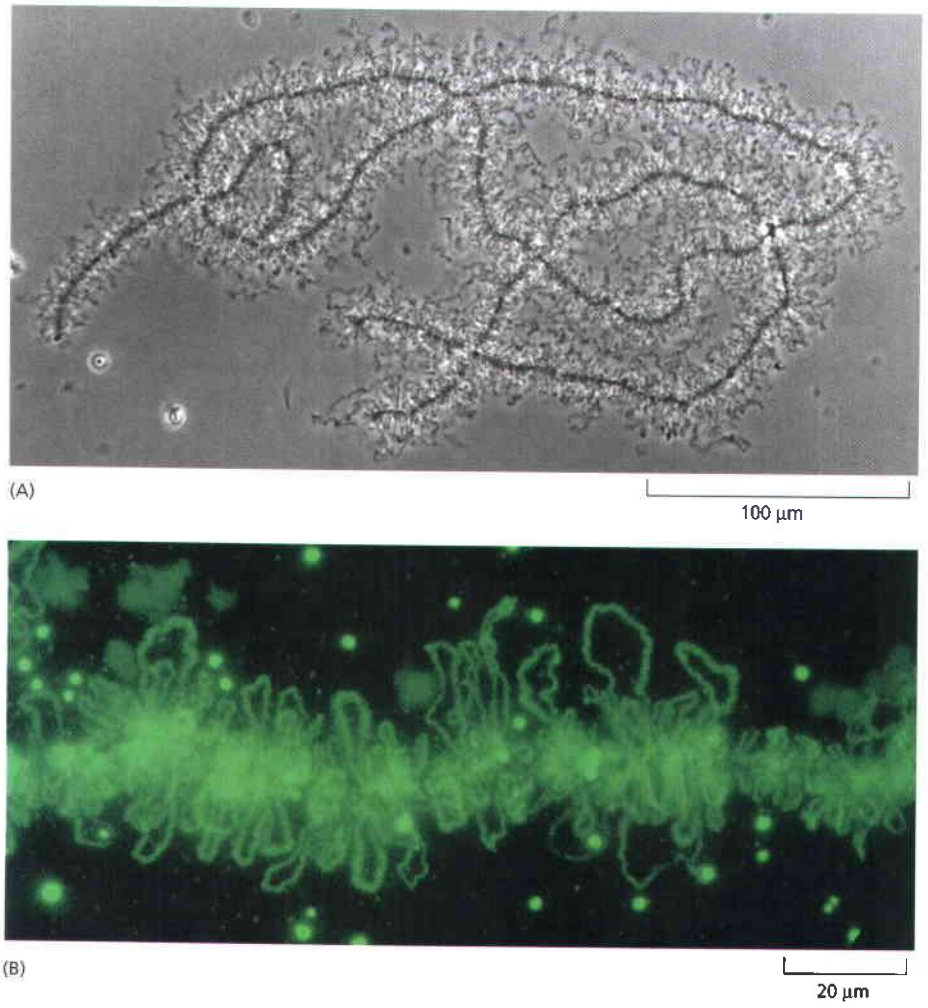


Figure 4–54 Lampbrush chromosomes. (A) A light micrograph of lampbrush chromosomes in an amphibian oocyte. Early in oocyte differentiation, each chromosome replicates to begin meiosis, and the homologous replicated chromosomes pair to form this highly extended structure containing a total of four replicated DNA molecules, or chromatids. The lampbrush chromosome stage persists for months or years, while the oocyte builds up a supply of materials required for its ultimate development into a new individual. (B) An enlarged region of a similar chromosome, stained with a fluorescent reagent that makes the loops active in RNA synthesis clearly visible. (Courtesy of Joseph G. Gall.)

provide a unique means for investigating some fundamental aspects of chromatin structure raised in the previous section. Next we describe how a typical interphase chromosome is arranged in the cell nucleus, focusing on human cells. Finally, we conclude by discussing the additional tenfold compaction that interphase chromosomes undergo during the process of mitosis.

Chromosomes Are Folded into Large Loops of Chromatin

Insight into the structure of the chromosomes in interphase cells has been obtained from studies of the stiff and extended meiotically paired chromosomes in growing amphibian oocytes (immature eggs). These very unusual **lampbrush chromosomes** (the largest chromosomes known) are clearly visible even in the light microscope, where they are seen to be organized into a series of large chromatin loops emanating from a linear chromosomal axis (**Figure 4–54**).

The organization of a lampbrush chromosome is illustrated in **Figure 4–55**. A given loop always contains the same DNA sequence, and it remains extended in the same manner as the oocyte grows. These chromosomes are producing large amounts of RNA for the oocyte, and most of the genes present in the DNA loops are being actively expressed. The majority of the DNA, however, is not in loops but remains highly condensed in the *chromomeres* on the axis, where genes are generally not expressed.

It is thought that the interphase chromosomes of all eucaryotes are similarly arranged in loops. Although these loops are normally too small and fragile to be easily observed in a light microscope, other methods can be used to infer their presence. For example, it has become possible to assess the frequency with

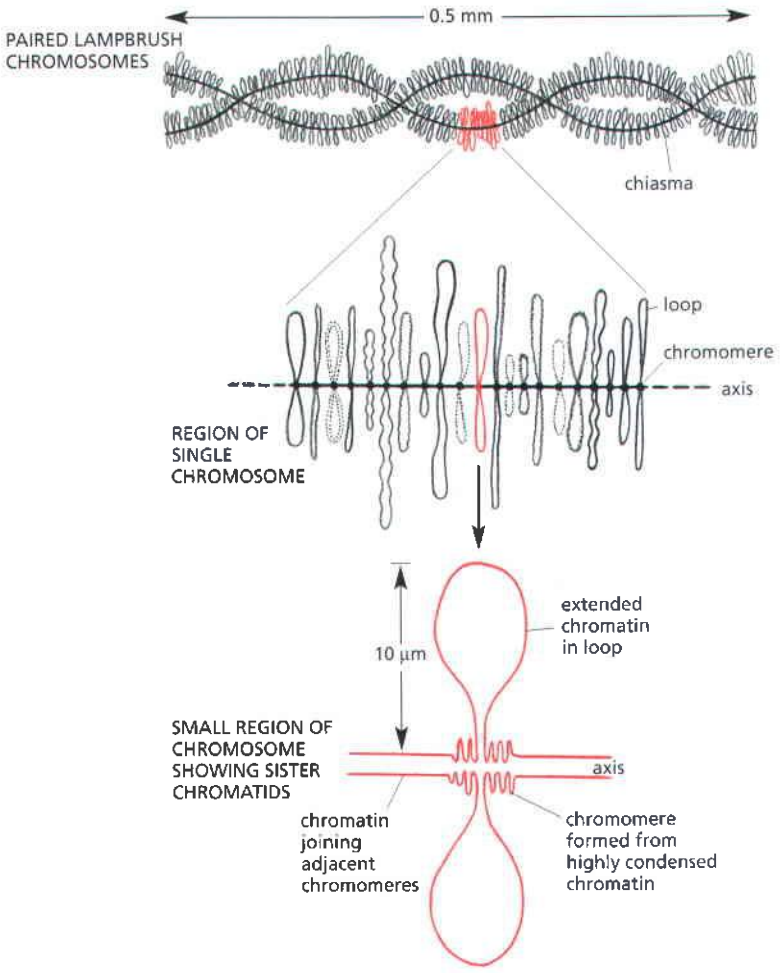


Figure 4–55 A model for the structure of a lampbrush chromosome. The set of lampbrush chromosomes in many amphibians contains a total of about 10,000 chromatin loops, although most of the DNA in each chromosome remains highly condensed in the chromomeres. Each loop corresponds to a particular DNA sequence. Four copies of each loop are present in each cell, since each of the two major units shown at the top consists of two closely apposed, newly replicated chromosomes. This four-stranded structure is characteristic of this stage of development of the oocyte, the diplotene stage of meiosis; see Figure 21–9.

which any two loci along an interphase chromosome are paired with each other, thus revealing likely candidates for the sites on chromatin that form the closely apposed bases of loop structures (**Figure 4–56**). These experiments and others suggest that the DNA in human chromosomes is organized into loops of

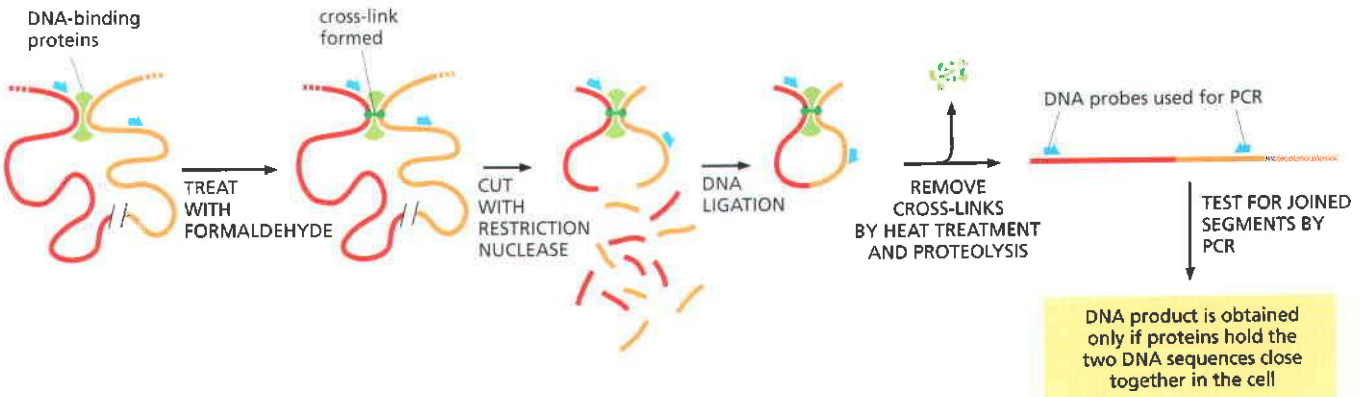


Figure 4–56 A method for determining the position of loops in interphase chromosomes. In this technique, known as the chromosome conformation capture (3C) method, cells are treated with formaldehyde to create the indicated covalent DNA–protein and DNA–DNA cross-links. The DNA is then treated with a restriction nuclease that chops the DNA into many pieces, cutting at strictly defined nucleotide sequences and forming sets of identical "cohesive ends" (see Figure 8–34). The cohesive ends can be made to join through their complementary base-pairing. Importantly, prior to the ligation step shown, the DNA is diluted so that the fragments that have been kept in close proximity to each other (through cross-linking) are the ones most likely to join. Finally, the cross-links are reversed and the newly ligated fragments of DNA are identified and quantified by PCR (the polymerase chain reaction, described in Chapter 8). By combining the frequency-of-association information generated by the 3C technique with DNA sequence information, structural models can be produced for the interphase conformation of chromosomes.

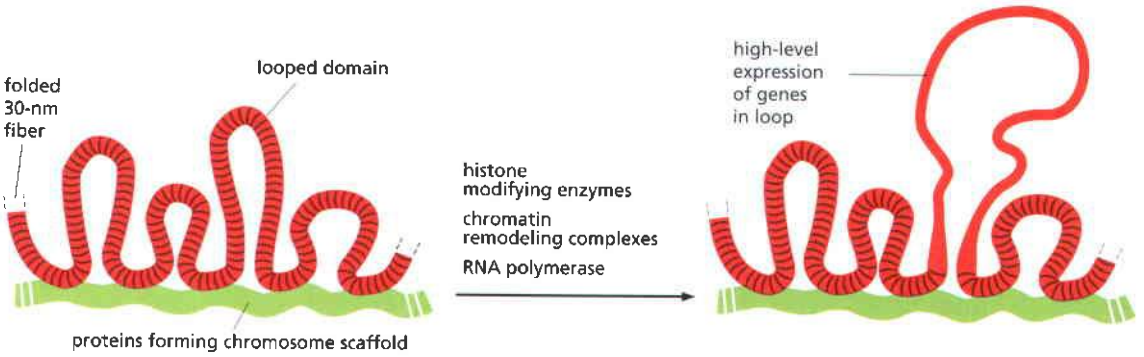


Figure 4–57 A model for the organization of an interphase chromosome. A section of an interphase chromosome is shown folded into a series of looped domains, each containing perhaps 50,000–200,000 nucleotide pairs of double-helical DNA condensed into a 30-nm fiber. The chromatin in each individual loop is further condensed through poorly understood folding processes that are reversed when the cell requires direct access to the DNA packaged in the loop. Neither the composition of the postulated chromosomal axis nor how the folded 30-nm fiber is anchored to it is clear. However, in mitotic chromosomes the bases of the chromosomal loops are enriched both in condensins and in DNA topoisomerase II enzymes, two proteins that may form much of the axis at metaphase (see Figure 4–74).

different lengths. A typical loop might contain between 50,000 and 200,000 nucleotide pairs of DNA, although loops of a million nucleotide pairs have also been suggested (**Figure 4–57**).

Polytene Chromosomes Are Uniquely Useful for Visualizing Chromatin Structures

Certain giant insect cells have grown to their enormous size through multiple cycles of DNA synthesis without cell division. Such cells with more than the normal DNA complement are said to be *polyploid* when they contain increased numbers of standard chromosomes. But in several types of secretory cells in fly larvae, all the homologous chromosome copies are held side by side, like drinking straws in a box, creating single large **polytene chromosomes**. Because polytene chromosomes can disperse to form a conventional polyploid cell in some cases, these two chromosomal states are closely related. The underlying structure of a polytene chromosome must therefore be similar to that of a normal chromosome.

Polyteny has been most studied in the larval salivary gland cells of the fruit fly *Drosophila*. When these polytene chromosomes are viewed in the light microscope, distinct alternating dark *bands* and light *interbands* are visible (**Figure 4–58**), each formed from a thousand identical DNA sequences arranged side-by-side in register. About 95% of the DNA in polytene chromosomes is in bands, and 5% is in interbands. A very thin band can contain 3000 nucleotide pairs, while a thick band may contain 200,000 nucleotide pairs in each of its chromatin strands. The chromatin in each band appears dark because the DNA is more condensed than the DNA in interbands; it may also contain a higher proportion of proteins (**Figure 4–59**).

There are approximately 3700 bands and 3700 interbands in the complete set of *Drosophila* polytene chromosomes. The bands can be recognized by their different thicknesses and spacings, and each one has been given a number to generate a chromosome "map" that has been indexed to the finished genome sequence of this fly.

The *Drosophila* polytene chromosomes provide a good starting point for examining how chromatin is organized on a large scale. In the previous section, we saw that there are many forms of chromatin, each of which contains nucleosomes with a different combination of modified histones. By reading this histone code, specific sets of non-histone proteins assemble on the nucleosomes to

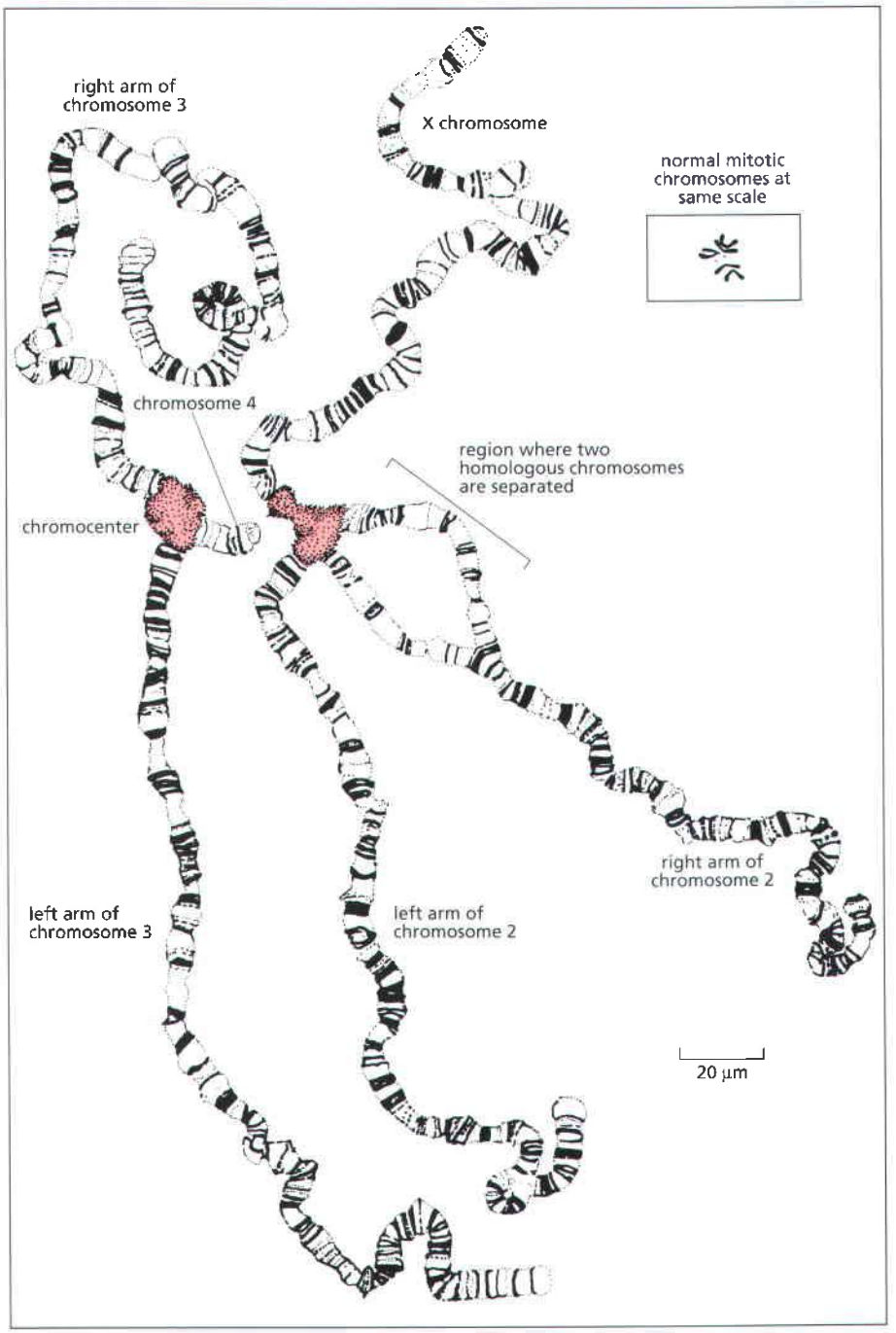


Figure 4–58 The entire set of polytene chromosomes in one *Drosophila* salivary cell. In this drawing of a light micrograph, the giant chromosomes have been spread out for viewing by squashing them against a microscope slide. *Drosophila* has four chromosomes, and there are four different chromosome pairs present. But each chromosome is tightly paired with its homolog (so that each pair appears as a single structure), which is not true in most nuclei (except in meiosis). Each chromosome has undergone multiple rounds of replication, and the homologues and all their duplicates have remained in exact register with each other, resulting in huge chromatin cables many DNA strands thick.

The four polytene chromosomes are normally linked together by heterochromatic regions near their centromeres that aggregate to create a single large chromocenter *(pink region)*. In this preparation, however, the chromocenter has been split into two halves by the squashing procedure used. (Adapted from T.S. Painter, *J. Hered.* 25:465–476, 1934. With permission from Oxford University Press.)

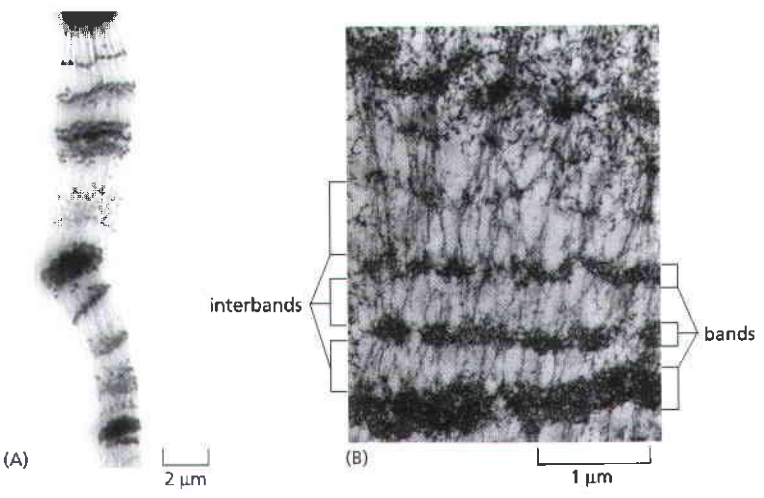


Figure 4–59 Micrographs of polytene chromosomes from *Drosophila* salivary glands. (A) Light micrograph of a portion of a chromosome. The DNA has been stained with a fluorescent dye, but a reverse image is presented here that renders the DNA *black* rather than *white*; the bands are clearly seen to be regions of increased DNA concentration. This chromosome has been processed by a high pressure treatment so as to show its distinct pattern of bands and interbands more clearly. (B) An electron micrograph of a small section of a *Drosophila* polytene chromosome seen in thin section. Bands of very different thickness can be readily distinguished, separated by interbands, which contain less condensed chromatin. (A, adapted from D.V. Novikov, I. Kireev and A.S. Belmont, *Nat. Methods* 4:483–485, 2007. With permission from Macmillan Publishers Ltd. B, courtesy of Veikko Sorsa.)

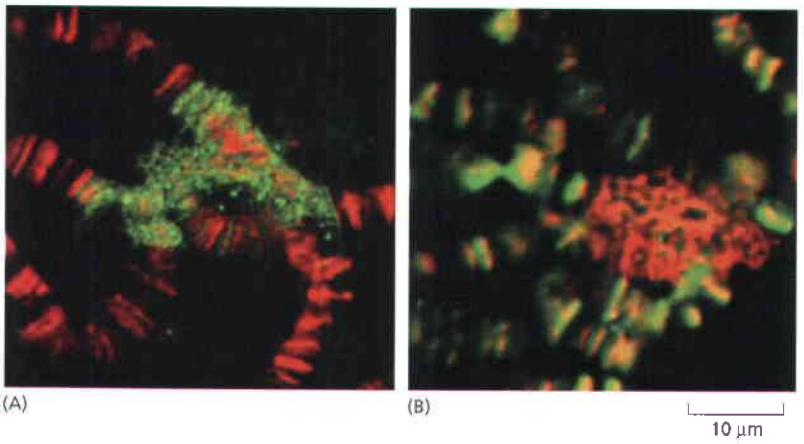


Figure 4–60 The pattern of histone modifications on *Drosophila* polytene chromosomes. Antibodies that specifically recognize different histone modifications can reveal where each modification is found with reference to the many bands and interbands on these chromosomes. In the two preparations shown here, the positions of two different markings on histone H3 tails are compared. In both cases, the antibody labeling the modified histone is *green*, and the DNA is stained *red*. Only a small region surrounding each chromocenter is shown. (A) Dimethyl Lys 9 *(green)* is a histone modification associated with the pericentric heterochromatin. It is seen to be associated with the chromocenter. (B) Acetylated Lys 9 *(green)* is a modification that is concentrated in histones associated with active genes. It is seen to be present in numerous bands on the chromosome arms, but not in the heterochromatic chromocenter. Similar experiments can be carried out to position many other modified histones, as well as the non-histone proteins (see, for example, Figure 22–45 for chromosomes stained for Polycomb). (Adapted from A. Ebert, S. Lein, G. Schotta and G. Reuter, *Chromosome Res.* 14:377–392, 2006. With permission from Springer.)

affect biological function in different ways. Some of these non-histone proteins can spread for long distances along the DNA, imparting a similar chromatin structure to contiguous regions of the genome (see Figure 4–46). Thus, in some regions, all of the chromatin has a similar structure and is separated from neighboring domains by barrier proteins (see Figure 4–47). At low resolution, the interphase chromosome can therefore be considered as a mosaic of chromatin structures, each containing particular nucleosome modifications associated with a particular set of non-histone proteins. (At a higher level of resolution one would also emphasize the many sequence-specific DNA-binding proteins that will be described in Chapter 7).

This view of an interphase chromosome helps us to interpret the results obtained from studies of polytene chromosomes. By staining with highly specific antibodies, one can show that differently modified histones (**Figure 4–60**), as well as distinct sets of non-histone proteins, are located on different polytene chromosome bands. This suggests a powerful general strategy. By employing combinations of antibodies that bind tightly to each of the many different histone modifications that create the histone code (see Figure 4–39), it may be possible to determine which combinations of modifications specify particular types of chromatin domains. And by carrying out similar experiments with antibodies that recognize each of the hundreds of different non-histone proteins in chromatin, one can attempt to decipher the many different meanings encoded in histone modifications.

There Are Multiple Forms of Heterochromatin

Molecular studies have led to a reevaluation of our view of heterochromatin. For many decades, heterochromatin was thought to be a single entity defined by its highly condensed structure and its ability to silence genes permanently. But if we define heterochromatin as a form of compact chromatin that can silence genes, be epigenetically inherited, and spread along chromosomes to cause position effect variegation (see Figure 4–36), it is clear that different types of heterochromatin exist. In fact, we have already considered three of these types in discussing the human centromere (see Figure 4–50).

Each domain of heterochromatin is thought to be formed by the cooperative assembly of a set of non-histone proteins. For example, classical pericentromeric heterochromatin contains more than six such proteins, including heterochromatin protein 1 (HP1), whereas the so-called Polycomb form of heterochromatin contains a similar number of proteins in a non-overlapping set (PcG proteins). There are hundreds of small blocks of heterochromatin spread across the arms of *Drosophila* polytene chromosomes, as identified by their late replication (discussed in Chapter 5). Antibody staining of these regions of heterochromatin suggests that the known forms of heterochromatin can account for no more than half of the heterchromatic polytene bands (**Figure 4–61**). Thus, other types of heterochromatin must exist whose protein composition is not known. It is likely

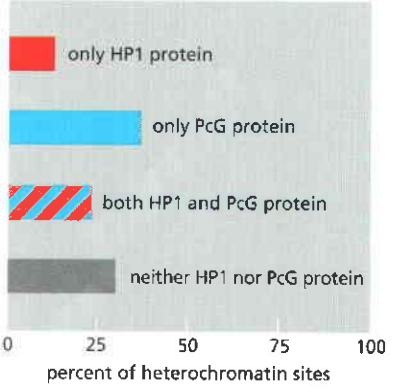


Figure 4–61 Evidence for multiple forms of heterochromatin. In this study, 240 late-replicating sites on the *Drosophila* polytene chromosome arms were examined for the presence of two non-histone proteins. These proteins are known to help compact two different forms of heterochromatin (see text). As indicated, antibody staining suggests that roughly half of the sites are packaged into forms of heterochromatin that are different from either of these two. Experiments such as these demonstrate that we have a great deal more to learn about the packaging of DNA in eucaryotes. (Data from I.F. Zhimulev and E.S. Belyaeva, *BioEssays* 25:1040–1051, 2003. With permission from John Wiley & Sons.)

that each of these types of heterochromatin is differently regulated and has different roles in the cell.

The chromatin structure in each domain ultimately depends on the proteins that bind to specific DNA sequences, and these are known to vary depending on the cell type and its stage of development in a multicellular organism. Thus, both the pattern of chromatin domains and their individual compositions (nucleosome modifications plus non-histone proteins) can vary between tissues. These differences make different genes accessible for genetic readout, helping to explain the cell diversification that accompanies embryonic development (described in Chapter 22). Comparisons of the polytene chromosomes in two different tissues of a fly lend support to this general idea: although the patterns of bands and interbands are largely the same, there are reproducible differences.

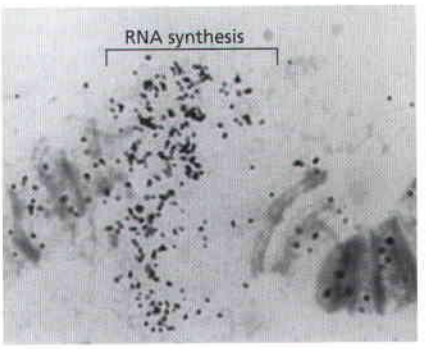


Figure 4–62 RNA synthesis in polytene chromosome puffs. An autoradiograph of a single puff in a polytene chromosome from the salivary glands of the freshwater midge *C. tentans*. As outlined in Chapter 1 and described in detail in Chapter 6, the first step in gene expression is the synthesis of an RNA molecule using the DNA as a template. The decondensed portion of the chromosome is undergoing RNA synthesis and has become labeled with ^{3}H-uridine (see p. 603), an RNA precursor molecule that is incorporated into growing RNA chains. (Courtesy of José Bonner.)

Chromatin Loops Decondense When the Genes Within Them Are Expressed

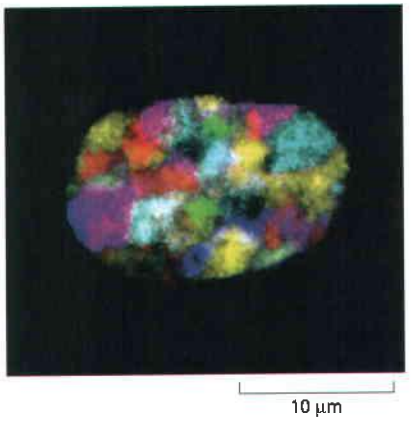
When an insect progresses from one developmental stage to another, distinctive *chromosome puffs* arise and old puffs recede in its polytene chromosomes as new genes become expressed and old ones are turned off (**Figure 4–62**). From inspection of each puff when it is relatively small and the banding pattern is still discernible, it seems that most puffs arise from the decondensation of a single chromosome band.

The individual chromatin fibers that make up a puff can be visualized with an electron microscope. In favorable cases, loops are seen, much like those observed in the amphibian lampbrush chromosomes discussed above. When not expressed, the loop of DNA assumes a thickened structure, possibly a folded 30-nm fiber, but when gene expression is occurring, the loop becomes more extended. In electron micrographs, the chromatin located on either side of the decondensed loop appears considerably more compact, suggesting that a loop constitutes a distinct functional domain of chromatin structure.

Observations made in human cells also suggest that highly folded loops of chromatin expand to occupy an increased volume when a gene within them is expressed. For example, quiescent chromosome regions from 0.4 to 2 million nucleotide pairs in length appear as compact dots in an interphase nucleus when visualized by fluorescence microscopy using FISH or other technologies. However, the same DNA is seen to occupy a larger territory when its genes are expressed, with elongated, punctate structures replacing the original dot.

Chromatin Can Move to Specific Sites Within the Nucleus to Alter Gene Expression

New ways of visualizing individual chromosomes have shown that each of the 46 interphase chromosomes in a human cell tends to occupy its own discrete territory within the nucleus (**Figure 4–63**). However, pictures such as these present only an average view of the DNA in each chromosome. Experiments that specifically localize the heterochromatic regions of a chromosome reveal that they are often closely associated with the nuclear lamina, regardless of the chromosome examined. And DNA probes that preferentially stain gene-rich regions of human



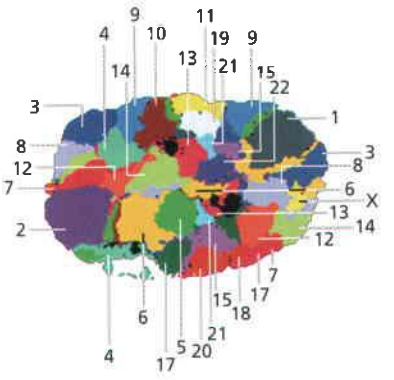


Figure 4–63 Simultaneous visualization of the chromosome territories for all of the human chromosomes in a single interphase nucleus. A FISH analysis using a different mixture of fluorochromes for marking the DNA of each chromosome, detected with seven color channels in a fluoresence microscope, allows each chromosome to be distinguished in three-dimensional reconstructions. Below the micrograph, each chromosome is identified in a schematic of the actual image. Note that the two homologous chromosomes (e.g., the two copies of chromosome 9), are not in general co-located. (From M.R. Speicher and N.P. Carter, *Nat. Rev. Genet.* 6:782–792, 2005. With permission from Macmillan Publishers Ltd.)

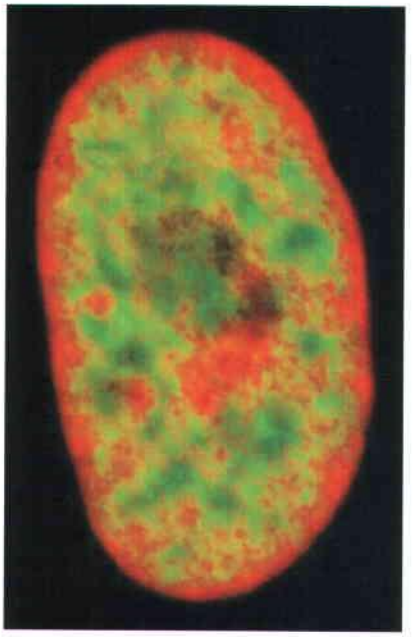


Figure 4–64 The distribution of gene-rich regions of the human genome in an interphase nucleus. Gene-rich regions have been visualized with a fluorescent probe that hybridizes to the Alu interspersed repeat, which is present in more than a million copies in the human genome (see Figure 5–75). For unknown reasons, these sequences cluster in chromosomal regions rich in genes. In this representation, regions enriched for the Alu sequence are *green,* regions depleted for these sequences are *red,* while the average regions are *yellow.* The gene-rich regions are seen to be depleted in the DNA near the nuclear envelope. (From A. Bolzer et al., *PLoS Biol.* 3:826–842, 2005. With permission from Public Library of Science.)

chromosomes produce a striking picture of the interphase nucleus that presumably reflects different average positions for active and inactive genes (**Figure 4–64**).

A variety of different types of experiments have led to the conclusion that the position of a gene in the interior of the nucleus changes when it becomes highly expressed. Thus, a region that becomes very actively transcribed is often found to extend out of its chromosome territory, as if in an extended loop (**Figure 4–65**). We will see in Chapter 6 that the initiation of transcription—the first step in gene expression—requires the assembly of over 100 proteins, and it makes sense that this would occur most rapidly in regions of the nucleus particularly rich in these proteins.

More generally, it is clear that the nucleus is very heterogeneous, with functionally different regions to which portions of chromosomes can move as they are subjected to different biochemical processes—such as when their gene expression changes (**Figure 4–66**). There is evidence that some of these nuclear regions are marked with different inositol phospholipids, reminiscent of the way that the same lipids are used to distinguish different membranes in the cytoplasm (see Figure 13–11). But what these lipids are attached to in the interior of the nucleus is a mystery, as the only known lipid-rich environments are the lipid bilayers of the nuclear envelope.

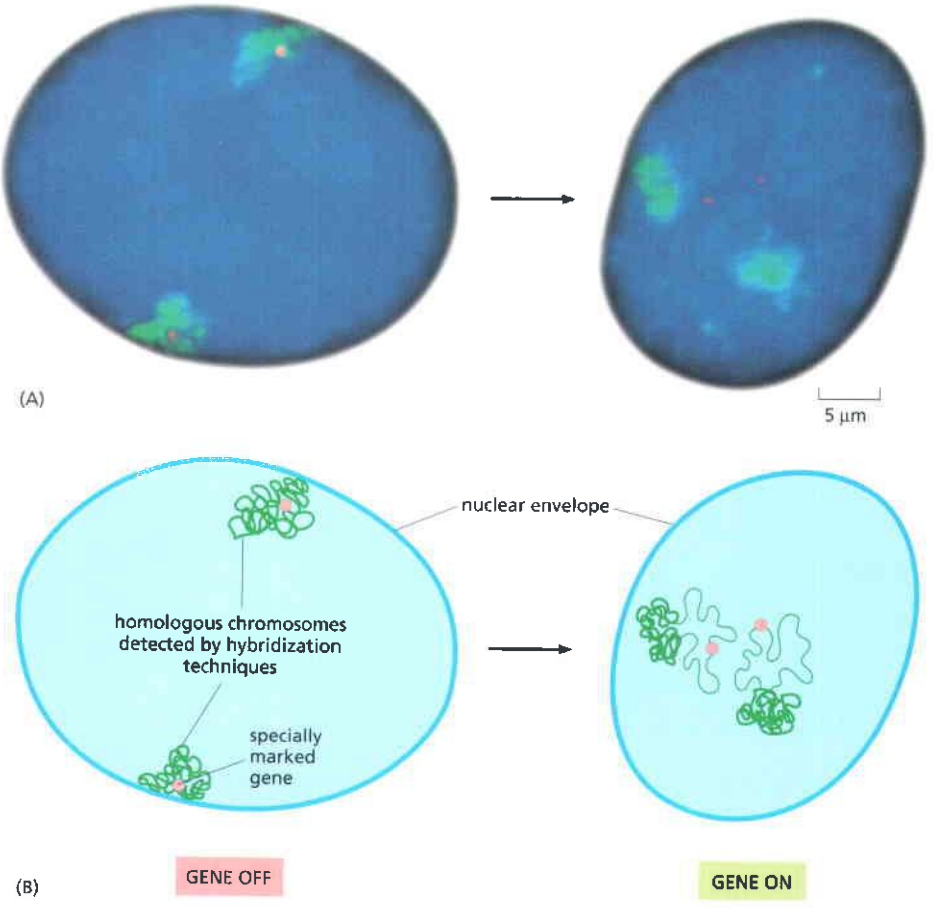


Figure 4–65 An effect of high levels of gene expression on the intranuclear location of chromatin. (A) Fluorescence micrographs of human nuclei showing how the position of a gene changes when it becomes highly transcribed. The region of the chromosome adjacent to the gene *(red)* is seen to leave its chromosomal territory *(green)* only when it is highly active. (B) Schematic representation of a large loop of chromatin that expands when the gene is on, and contracts when the gene is off. Other genes that are less actively expressed can be shown by the same methods to remain inside their chromosomal territory when transcribed. (From J.R. Chubb and W.A. Bickmore, *Cell* 112:403–406, 2003. With permission from Elsevier.)

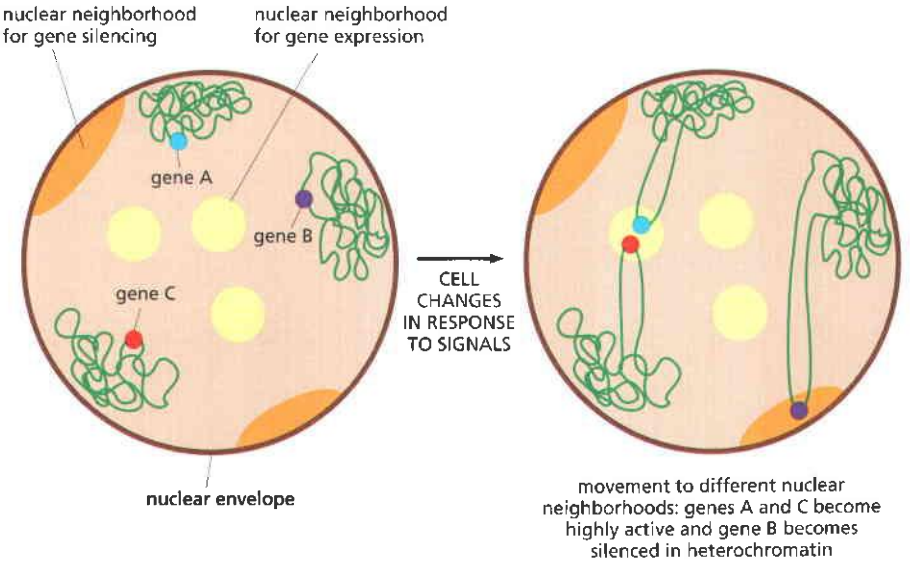


Figure 4–66 The movement of genes to different regions of the nucleus when their expression changes. The interior of the nucleus is very heterogeneous, and different nuclear neighborhoods are known to have distinct effects on gene expression. Movements such as those indicated here presumably reflect changes in the binding affinities that the chromatin and RNA molecules surrounding a gene have for different nuclear neighborhoods. It is thought that the movement is driven by diffusion and does not require a directed movement process, inasmuch as each region of a chromosome can be seen to undergo constant random motion when marked in a way that allows its position to be followed in a living cell.

Networks of Macromolecules Form a Set of Distinct Biochemical Environments inside the Nucleus

In Chapter 6, we describe the function of a variety of subcompartments that are present within the nucleus. The largest and most obvious of these is the nucleolus, a structure well known to microscopists even in the 19th century (see Figure 4–9). Nucleolar regions consist of networks of RNAs and proteins surrounding transcribing ribosomal RNA genes, often existing as multiple nucleoli. The nucleolus is the cell's site of ribosome assembly and maturation, as well as the place where many other specialized reactions occur.

A variety of less obvious organelles are also present inside the nucleus. For example, spherical structures called Cajal bodies and interchromatin granule clusters are present in most plant and animal cells (**Figure 4–67**). Like the nucleolus, these organelles are composed of selected protein and RNA molecules that bind together to create networks that are highly permeable to other protein and RNA molecules in the surrounding nucleoplasm (**Figure 4–68**).

Structures such as these can create distinct biochemical environments by immobilizing select groups of macromolecules, as can other networks of proteins and RNA molecules associated with nuclear pores and with the nuclear envelope. In principle, this allows the molecules that enter these spaces to be processed with great efficiency through complex reaction pathways. Highly permeable, fibrous networks of this sort can thereby impart many of the kinetic advantages of compartmentalization (see p. 186) to reactions that take place in the nucleus (**Figure 4–69**A). However, unlike the membrane-bound compartments in the cytoplasm (discussed in Chapter 12), these nuclear subcompartments—lacking a lipid bilayer membrane—can neither concentrate nor exclude specific small molecules.

The cell has a remarkable ability to construct distinct biochemical environments inside the nucleus. Those thus far mentioned facilitate various aspects of gene expression to be discussed in Chapter 6 (see Figure 6–49). Like the nucleolus, these subcompartments appear to form only as needed, and they create a high local concentration of the many different enzymes and RNA molecules needed for a particular process. In an analogous way, when DNA is damaged by irradiation, the set of enzymes needed to carry out DNA repair are observed to congregate in discrete foci inside the nucleus, creating "repair factories" (see Figure 5–60). And nuclei often contain hundreds of discrete foci representing factories for DNA or RNA synthesis.

It seems likely that all of these entities make use of the type of tethering illustrated in Figure 4–69B, where long flexible lengths of polypeptide chain (or some other polymer) are interspersed with binding sites that concentrate the multiple proteins and/or RNA molecules that are needed to catalyze a particular process. Not surprisingly, tethers are similarly used to help to speed biological processes

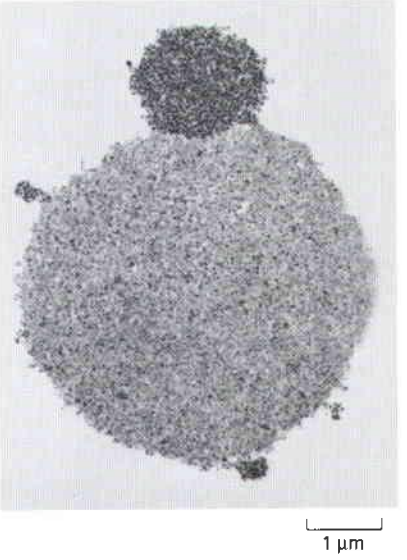


Figure 4–67 Electron micrograph showing two very common fibrous nuclear subcompartments. The large sphere here is a Cajal body. The smaller darker sphere is an interchromatin granule cluster, also known as a spreckle (see also Figure 6–49). These "subnuclear organelles" are from the nucleus of a *Xenopus* oocyte. (From K.E. Handwerger and J.G. Gall, *Trends Cell Biol.* 16:19–26, 2006. With permission from Elsevier.)

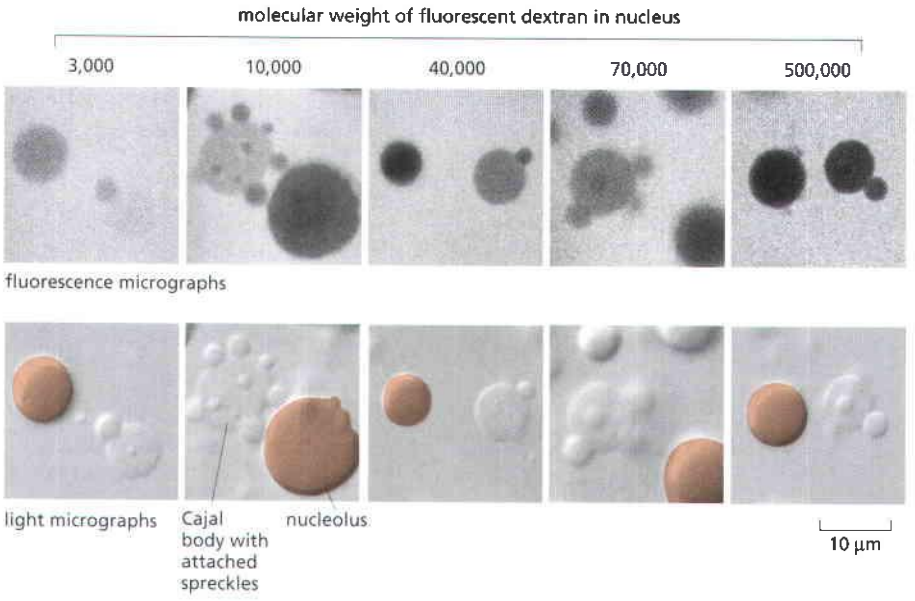


Figure 4–68 An experiment showing that the subnuclear organelles are highly permeable to macromolecules. In these micrographs of a living oocyte nucleus, the top row compares the fluorescence of the interiors of nucleoli, Cajal bodies, and spreckles to the fluorescence of the surrounding nucleoplasm, 12 hours after fluorescent dextrans of the indicated molecular weight had been injected into the nucleoplasm. The brightness of each organelle reflects its permeability, with the most permeable organelle being the brightest. For comparison, the bottom row presents normal light micrographs of the same microscope fields, with the nucleolus in each field of view marked *brown* to distinguish it. Cajal bodies can be seen to be more permeable than nucleoli. However, quantitation shows that a great deal of dextran enters each organelle, even for the largest dextran tested. (From K.E. Handwerger, J.A. Cordero and J.G. Gall, *Mol. Biol. Cell* 16:202–211, 2005. With permission from American Society of Cell Biology.)

in the cytoplasm, increasing specific reaction rates (for example, see Figure 16–38).

Is there also is an intranuclear framework, analogous to the cytoskeleton, on which chromosomes and other components of the nucleus are organized? The *nuclear matrix*, or *scaffold*, has been defined as the insoluble material left in the nucleus after a series of biochemical extraction steps. Many of the proteins and RNA molecules that form this insoluble material are likely to be derived from the fibrous subcompartments of the nucleus just discussed, while others seem to be proteins that help to form the base of chromosomal loops or to attach chromosomes to other structures in the nucleus. Whether or not the nucleus also contains

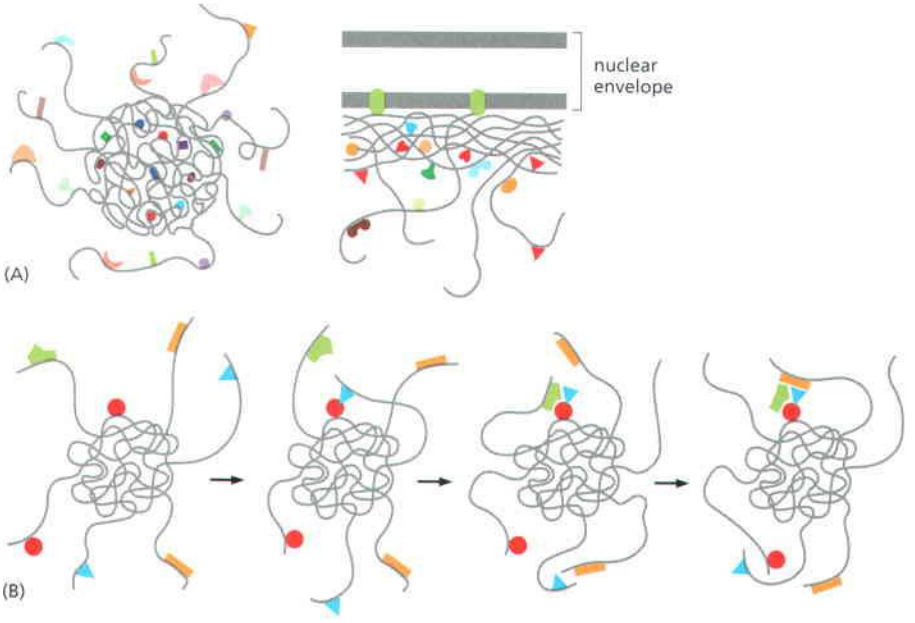


Figure 4–69 Effective compartmentalization without a bilayer membrane. (A) Schematic illustration of the organization of a spherical subnuclear organelle (*left*) and of a postulated similarly organized subcompartment just beneath the nuclear envelope (*right*). In both cases, RNAs and/or proteins (*gray*) associate to form highly porous, gel-like structures that contain binding sites for other specific proteins and RNA molecules (*colored objects*). (B) How the tethering of a selected set of proteins and RNA molecules to long flexible polymer chains, as in A, could create "staging areas" that greatly speed the rates of reactions in subcompartments of the nucleus. The reactions catalyzed will depend on the particular macromolecules that are localized by the tethering. The same type of rate accelerations are of course expected for similar subcompartments established elsewhere in the cell (see also Figure 3–80C).

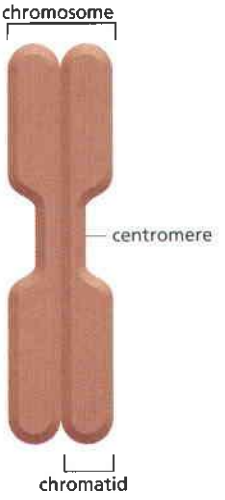


Figure 4–70 A typical mitotic chromosome at metaphase. Each sister chromatid contains one of two identical daughter DNA molecules generated earlier in the cell cycle by DNA replication (see also Figure 17–26).

long filaments that form organized tracks on which nuclear components can move, analogous to some of the filaments in the cytoplasm, is still disputed.

Mitotic Chromosomes Are Formed from Chromatin in Its Most Condensed State

Having discussed the dynamic structure of interphase chromosomes, we now turn to mitotic chromosomes. The chromosomes from nearly all eucaryotic cells become readily visible by light microscopy during mitosis, when they coil up to form highly condensed structures. This condensation reduces the length of a typical interphase chromosome only about tenfold, but it produces a dramatic change in chromosome appearance.

Figure 4–70 depicts a typical **mitotic chromosome** at the metaphase stage of mitosis (for the stages of mitosis, see Figure 17–3). The two daughter DNA molecules produced by DNA replication during interphase of the cell-division cycle are separately folded to produce two sister chromosomes, or *sister chromatids*, held together at their centromeres (see also Figure 4–50). These chromosomes are normally covered with a variety of molecules, including large amounts of RNA–protein complexes. Once this covering has been stripped away, each chromatid can be seen in electron micrographs to be organized into loops of chromatin emanating from a central scaffolding (**Figure 4–71**). Experiments using DNA hybridization to detect specific DNA sequences demonstrate that the order of visible features along a mitotic chromosome at least roughly reflects the order of genes along the DNA molecule. Mitotic chromosome condensation can thus be thought of as the final level in the hierarchy of chromosome packaging (**Figure 4–72**).

The compaction of chromosomes during mitosis is a highly organized and dynamic process that serves at least two important purposes. First, when condensation is complete (in metaphase), sister chromatids have been disentangled from each other and lie side by side. Thus, the sister chromatids can easily separate when the mitotic apparatus begins pulling them apart. Second, the compaction of chromosomes protects the relatively fragile DNA molecules from being broken as they are pulled to separate daughter cells.

The condensation of interphase chromosomes into mitotic chromosomes begins in early M phase, and it is intimately connected with the progression of the cell cycle, as discussed in detail in Chapter 17. During M phase, gene expression shuts down, and specific modifications are made to histones that help to reorganize the chromatin as it compacts. The compaction is aided by a class of proteins called *condensins*, which use the energy of ATP hydrolysis to help coil the two DNA molecules in an interphase chromosome to produce the two chromatids of a mitotic chromosome. Condensins are large protein complexes built from SMC protein dimers: these dimers form when two stiff, elongated protein monomers join at their tails to form a hinge, leaving two globular head domains at the other end that bind DNA and hydrolyze ATP (**Figure 4–73**). When added to purified DNA, condensins can make large right-handed loops in DNA molecules in a reaction that requires ATP. Although it is not yet known how they act on chromatin, the coiling model shown in Figure 4–73C is based on the fact that condensins are a major structural component that end up at the core of metaphase chromosomes, with about one molecule of condensin for every

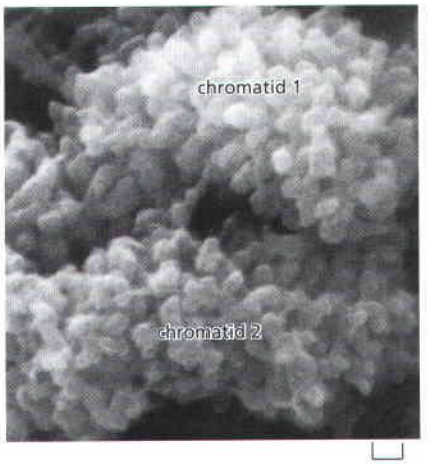


Figure 4–71 A scanning electron micrograph of a region near one end of a typical mitotic chromosome. Each knoblike projection is believed to represent the tip of a separate looped domain. Note that the two identical paired chromatids (drawn in Figure 4–70) can be clearly distinguished. (From M.P. Marsden and U.K. Laemmli, *Cell* 17:849–858, 1979. With permission from Elsevier.)

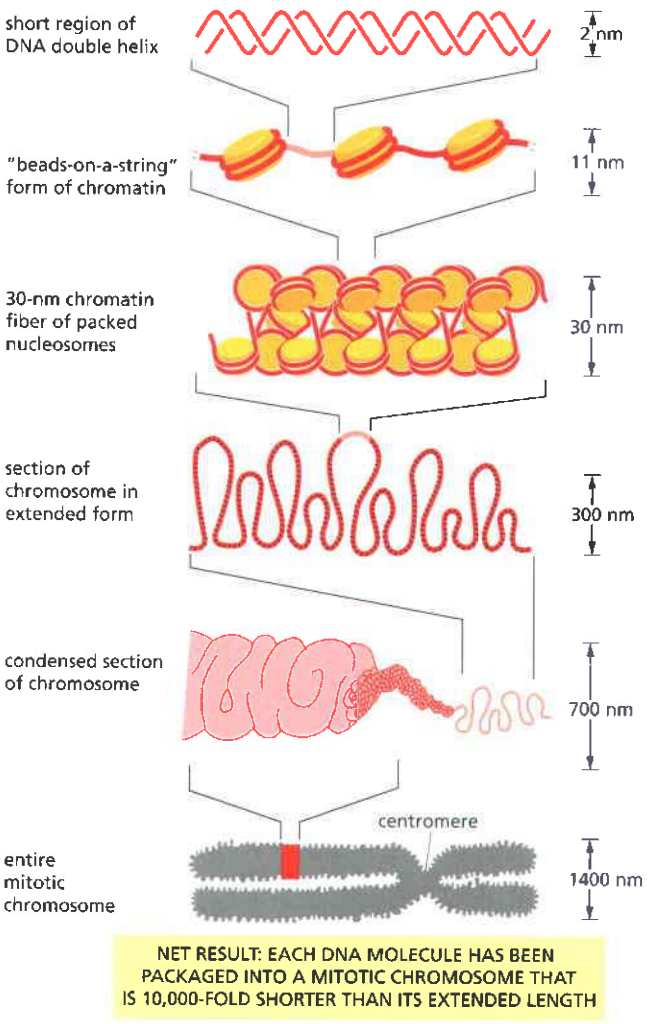
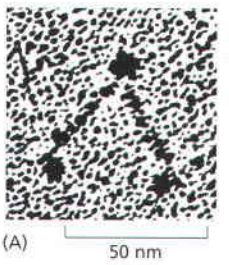


Figure 4–72 Chromatin packing. This model shows some of the many levels of chromatin packing postulated to give rise to the highly condensed mitotic chromosome.



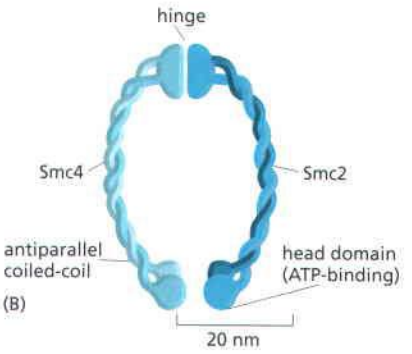
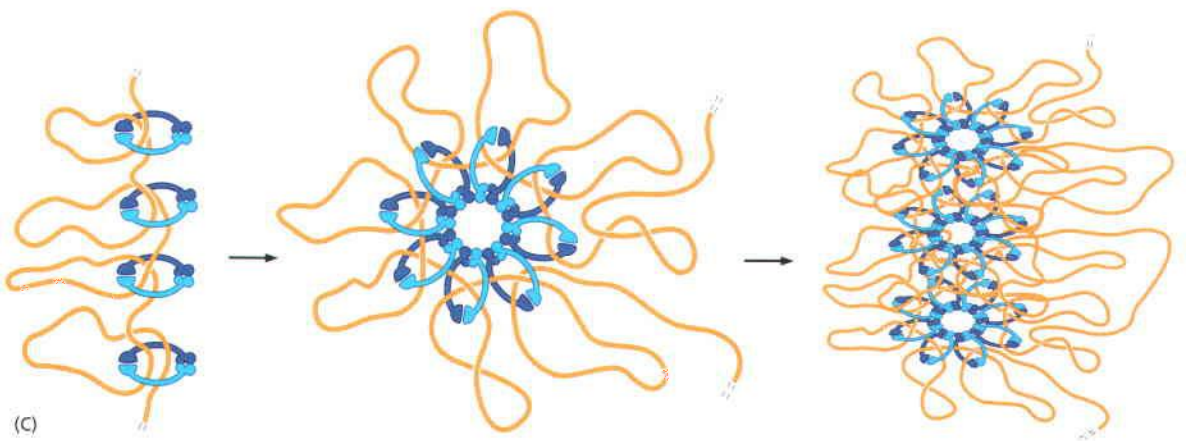


Figure 4–73 The SMC proteins in condensins. (A) Electron micrographs of a purified SMC dimer. (B) The structure of a SMC dimer. The long central region of this protein is an antiparallel coiled-coil (see Figure 3–9) with a flexible hinge in its middle. (C) A model for the way in which the SMC proteins in condensins might compact chromatin. In reality, SMC proteins are components of a much larger condensin complex. It has been proposed that, in the cell, condensins coil long strings of looped chromatin domains (see Figure 4–57). In this way, the condensins could form a structural framework that maintains the DNA in a highly organized state during metaphase of the cell cycle. (A, courtesy of H.P. Erickson; B and C, adapted from T. Hirano, *Nat. Rev. Mol. Cell Biol.* 7:311–322, 2006. With permission from Macmillan Publishers Ltd.)



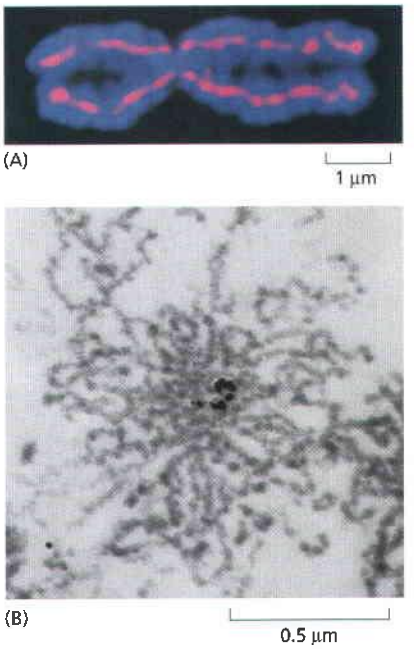


Figure 4–74 The location of condensin in condensed mitotic chromosomes. (A) Fluorescence micrograph of a human chromosome at mitosis, stained with an antibody that localizes condensin. In chromosomes that are this highly condensed, the condensin is seen to be concentrated in punctate structures along the chromosome axis. Similar experiments show a similar location for DNA topoisomerase II, an enzyme that makes reversible double-strand breaks in DNA that allow one DNA double helix to pass through another (see Figure 5–23). (B) Immunogold electron microscopy reveals localization of condensin (*black* dots). Here a chromatid is seen in cross section, with the chromosome axis perpendicular to the plane of the paper. (A, from K. Maeshima and U.K. Laemmli, *Dev. Cell* 4:467–480, 2003. With permission from Elsevier. B, courtesy of U.K. Laemmli, from K. Maeshima, M. Eltsov and U.K. Laemmli, *Chromosoma* 114:365–375, 2005. With permission from Springer.)

10,000 nucleotides of DNA (**Figure 4–74**). When condensins are experimentally depleted from a cell, chromosome condensation still occurs, but the process is abnormal.

Summary

Chromosomes are generally decondensed during interphase, so that the details of their structure are difficult to visualize. Notable exceptions are the specialized lampbrush chromosomes of vertebrate oocytes and the polytene chromosomes in the giant secretory cells of insects. Studies of these two types of interphase chromosomes suggest that each long DNA molecule in a chromosome is divided into a large number of discrete domains organized as loops of chromatin, each loop probably consisting of a 30-nm chromatin fiber that is compacted by further folding. When genes contained in a loop are expressed, the loop unfolds and allows the cell's machinery access to the DNA.

Interphase chromosomes occupy discrete territories in the cell nucleus; that is, they are not extensively intertwined. Euchromatin makes up most of interphase chromosomes and, when not being transcribed, it probably exists as tightly folded 30-nm fibers. However, euchromatin is interrupted by stretches of heterochromatin, in which the 30-nm fibers are subjected to additional packing that usually renders it resistant to gene expression. Heterochromatin exists in several forms, some of which are found in large blocks in and around centromeres and near telomeres. But heterochromatin is also present at many other positions on chromosomes, where it can serve to regulate developmentally important genes.

The interior of the nucleus is highly dynamic, with heterochromatin often positioned near the nuclear envelope and loops of chromatin moving away from their chromosome territory when genes are very highly expressed. This reflects the existence of nuclear subcompartments, where different sets of biochemical reactions are facilitated by an increased concentration of selected proteins and RNAs. The components involved in forming a subcompartment can self-assemble into discrete organelles such as nucleoli or Cajal bodies; they can also be tethered to fixed structures such as the nuclear envelope.

During mitosis, gene expression shuts down and all chromosomes adopt a highly condensed conformation in a process that begins early in M phase to package the two DNA molecules of each replicated chromosome as two separately folded chromatids. This process is accompanied by histone modifications that facilitate chromatin packing. However, satisfactory completion of this orderly process, which reduces the end-to-end distance of each DNA molecule from its interphase length by an additional factor of ten, requires condensin proteins.

HOW GENOMES EVOLVE

In this chapter, we have discussed the structure of genes and the ways that they are packaged and arranged in chromosomes. In this final section, we provide an overview of some of the ways that genes and genomes have evolved over time to produce the vast diversity of modern-day life forms on our planet. Genome

sequencing has revolutionized our view of the process of molecular evolution, uncovering an astonishing wealth of information about specific family relationships among organisms, as well as illuminating evolutionary mechanisms more generally.

It is perhaps not surprising that genes with similar functions can be found in a diverse range of living things. But the great revelation of the past 25 years has been the discovery that the actual nucleotide sequences of many genes are sufficiently well conserved that **homologous** genes—that is, genes that are similar in both their nucleotide sequence and function because of a common ancestry—can often be recognized across vast phylogenetic distances. For example, unmistakable homologs of many human genes are easy to detect in such organisms as nematode worms, fruit flies, yeasts, and even bacteria. In many cases, the resemblance is so close that the protein-coding portion of a yeast gene can be substituted with its human homolog—even though we and yeast are separated by more than a billion years of evolutionary history.

As emphasized in Chapter 3, the recognition of sequence similarity has become a major tool for inferring gene and protein function. Although finding a sequence match does not guarantee similarity in function, it has proven to be an excellent clue. Thus, it is often possible to predict the function of genes in humans for which no biochemical or genetic information is available simply by comparing their nucleotide sequences with the sequences of genes in other organisms.

In general, gene sequences are more tightly conserved than is overall genome structure. As we saw earlier, other features of genome organization such as genome size, number of chromosomes, order of genes along chromosomes, abundance and size of introns, and amount of repetitive DNA are found to differ greatly among organisms, as does the number of genes that an organism contains.

The number of genes is only very roughly correlated with the phenotypic complexity of an organism (see Table 1–1). Much of the increase in gene number observed with increasing biological complexity involves the expansion of families of closely related genes, an observation that establishes gene duplication and divergence as major evolutionary processes. Indeed, it is likely that all present-day genes are descendants—via the processes of duplication, divergence, and reassortment of gene segments—of a few ancestral genes that existed in early life forms.

Genome Alterations are Caused by Failures of the Normal Mechanisms for Copying and Maintaining DNA

Cells in the germline do not have specialized mechanisms for creating changes in the structures of their genomes: evolution depends instead on accidents and mistakes followed by nonrandom survival. Most of the genetic changes that occur result simply from failures in the normal mechanisms by which genomes are copied or repaired when damaged, although the movement of transposable DNA elements also plays an important role. As we will discuss in Chapter 5, the mechanisms that maintain DNA sequences are remarkably precise—but they are not perfect. For example, because of the elaborate DNA-replication and DNA-repair mechanisms that enable DNA sequences to be inherited with extraordinary fidelity, along a given line of descent only about one nucleotide pair in a thousand is randomly changed in the germline every million years. Even so, in a population of 10,000 diploid individuals, every possible nucleotide substitution will have been "tried out" on about 20 occasions in the course of a million years—a short span of time in relation to the evolution of species.

Errors in DNA replication, DNA recombination, or DNA repair can lead either to simple changes in DNA sequence—such as the substitution of one base pair for another—or to large-scale genome rearrangements such as deletions, duplications, inversions, and translocations of DNA from one chromosome to another. In addition to these failures of the genetic machinery, the various mobile DNA elements that will be described in Chapter 5 are an important source of genomic change (see Table 5–3, p. 318). These transposable DNA elements (*transposons*)

are parasitic DNA sequences that colonize genomes and can spread within them. In the process, they often disrupt the function or alter the regulation of existing genes. On occasion, they can even create altogether novel genes through fusions between transposon sequences and segments of existing genes. Over long periods of evolutionary time, transposons have profoundly affected the structure of genomes. In fact, nearly half of the DNA in the human genome has recognizable sequence similarity with known transposon sequences, thereby indicating that these sequences are remnants of past transposition events (see Figure 4–17). Even more of our genome is no doubt derived from transposition events that occurred so long ago (>10^8 years) that the sequences can no longer be traced to transposons.

The Genome Sequences of Two Species Differ in Proportion to the Length of Time That They Have Separately Evolved

The differences between the genomes of species alive today have accumulated over more than 3 billion years. Lacking a direct record of changes over time, we can nevertheless reconstruct the process of genome evolution from detailed comparisons of the genomes of contemporary organisms.

The basic tool of comparative genomics is the phylogenetic tree. A simple example is the tree describing the divergence of humans from the great apes (**Figure 4–75**). The primary support for this tree comes from comparisons of gene or protein sequences. For example, comparisons between the sequences of human genes or proteins and those of the great apes typically reveal the fewest differences between human and chimpanzee and the most between human and orangutan.

For closely related organisms such as humans and chimpanzees, it is relatively easy to reconstruct the gene sequences of the extinct, last common ancestor of the two species (**Figure 4–76**). The close similarity between human and chimpanzee genes is mainly due to the short time that has been available for the accumulation of mutations in the two diverging lineages, rather than to functional constraints that have kept the sequences the same. Evidence for this view comes from the observation that even DNA sequences whose nucleotide order is functionally unconstrained—such as the sequences that code for the fibrinopeptides (see p. 264) or the third position of "synonymous" codons (codons specifying the same amino acid—see Figure 4–76)—are nearly identical in humans and chimpanzees.

For much less closely related organisms, such as humans and chickens (which have evolved separately for about 300 million years), the sequence conservation found in genes is largely due to **purifying selection** (that is, selection that eliminates individuals carrying mutations that interfere with important genetic functions), rather than to an inadequate time for mutations to occur. As a result, protein-coding, RNA-coding, and regulatory sequences in the DNA are often remarkably conserved. In contrast, most DNA sequences in the human and chicken genomes have diverged so far due to multiple mutations that it is often impossible to align them with one another.

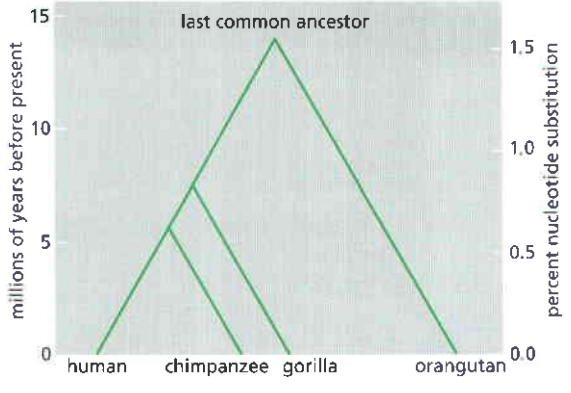


Figure 4–75 A phylogenetic tree showing the relationship between the human and the great apes based on nucleotide sequence data. As indicated, the sequences of the genomes of all four species are estimated to differ from the sequence of the genome of their last common ancestor by a little over 1.5%. Because changes occur independently on both diverging lineages, pairwise comparisons reveal twice the sequence divergence from the last common ancestor. For example, human–orangutan comparisons typically show sequence divergences of a little over 3%, while human–chimpanzee comparisons show divergences of approximately 1.2%. (Modified from F.C. Chen and W.H. Li, *Am. J. Hum. Genet.* 68:444–456, 2001. With permission from University of Chicago Press.)

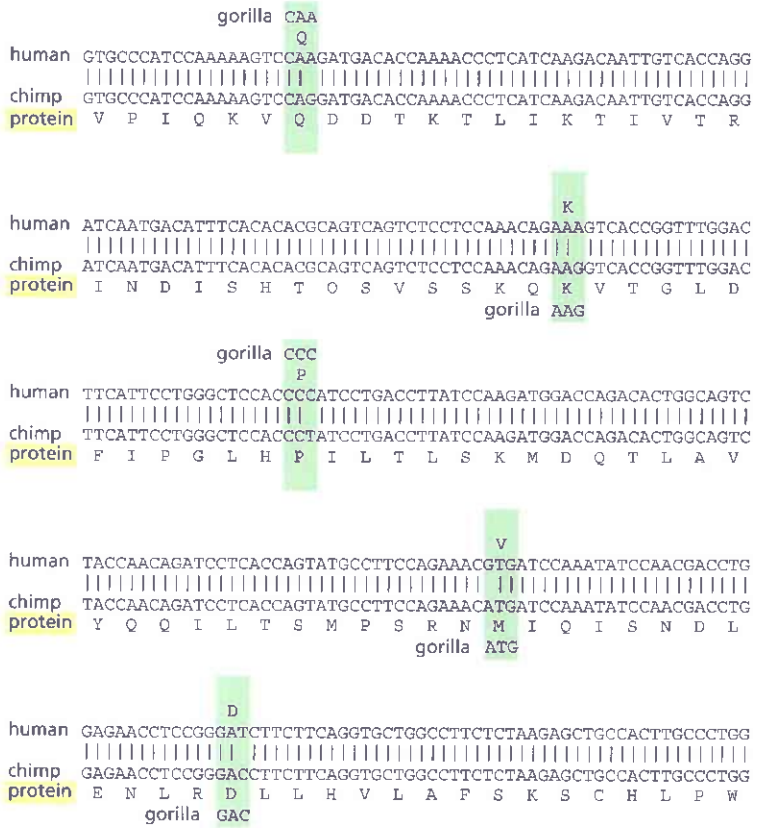


Figure 4–76 Tracing the ancestral sequence from a sequence comparison of the coding regions of human and chimpanzee leptin genes. Leptin is a hormone that regulates food intake and energy utilization in response to the adequacy of fat reserves. As indicated by the codons boxed in green, only 5 nucleotides (of 441 total) differ between these two sequences. Moreover, when the amino acids encoded by both the human and chimpanzee sequences are examined, in only one of the 5 positions does the encoded amino acid differ. For each of the 5 variant nucleotide positions, the corresponding sequence in the gorilla is also indicated. In two cases, the gorilla sequence agrees with the human sequence, while in three cases it agrees with the chimpanzee sequence.

What was the sequence of the leptin gene in the last common ancestor? An evolutionary model that seeks to minimize the number of mutations postulated to have occurred during the evolution of the human and chimpanzee genes would assume that the leptin sequence of the last common ancestor was the same as the human and chimpanzee sequences when they agree; when they disagree, it would use the gorilla sequence as a tie-breaker. For convenience, only the first 300 nucleotides of the leptin coding sequences are given. The remaining 141 are identical between humans and chimpanzees.

Phylogenetic Trees Constructed from a Comparison of DNA Sequences Trace the Relationships of All Organisms

Integration of phylogenetic trees based on molecular sequence comparisons with the fossil record has led to the best available view of the evolution of modern life forms. The fossil record remains important as a source of absolute dates based on the decay of radioisotopes in the rock formations in which fossils are found. However, precise divergence times between species are difficult to establish from the fossil record, even for species that leave good fossils with distinctive morphology.

Such integrated phylogenetic trees suggest that changes in the sequences of particular genes or proteins tend to occur at a nearly constant rate, although rates that differ from the norm by as much as twofold are observed in particular lineages. As discussed above and in Chapter 5, this "molecular clock" runs most rapidly and regularly in sequences that are not subject to purifying selection—such as intergenic regions, portions of introns that lack splicing or regulatory signals, and genes that have been irreversibly inactivated by mutation (the so-called pseudogenes). The clock runs most slowly for sequences that are subject to strong functional constraints—for example, the amino acid sequences of proteins such as actin that engage in specific interactions with large numbers of other proteins and whose structure is therefore highly constrained (see, for example, Figure 16–18).

Occasionally, rapid change is seen in a previously highly conserved sequence. As discussed later in this chapter, such episodes are especially interesting because they are thought to reflect periods of strong positive selection for mutations that conferred a selective advantage in the particular lineage where the rapid change occurred.

Molecular clocks run at rates that are determined both by mutation rates and by the degree of purifying selection on particular sequences. Therefore, a completely different calibration is required for those genes replicated and repaired by different systems within cells. Most notably, in animals, although not in plants, clocks based on functionally unconstrained mitochondrial DNA sequences run

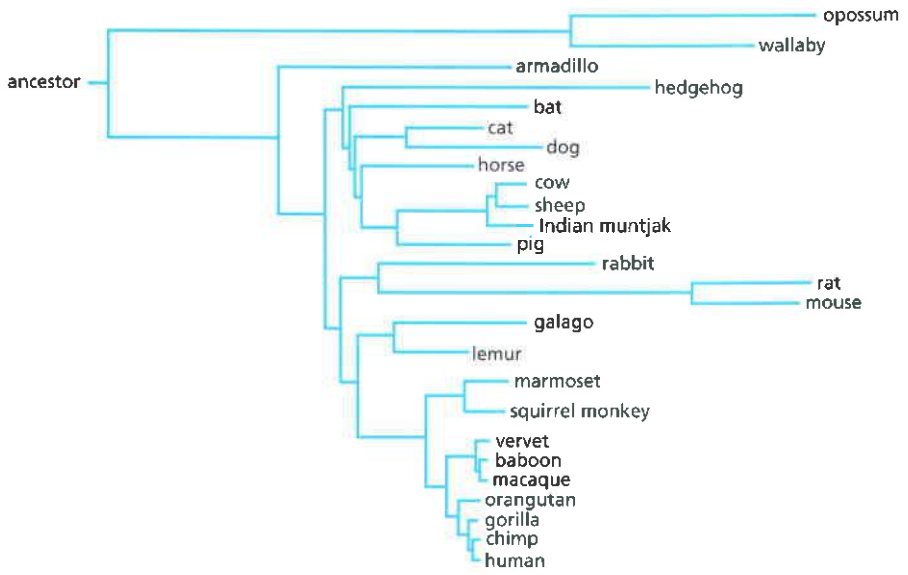


Figure 4–77 A phylogenetic tree highlighting some of the mammals whose genomes are being extensively studied. The length of each line is proportional to the number of "neutral substitutions"—representing the nucleotide changes observed in the absence of purifying selection. (Adapted from G.M. Cooper et al., *Genome Res.* 15:901–913, 2005. With permission from Cold Spring Harbor Laboratory Press.)

much faster than clocks based on functionally unconstrained nuclear sequences, due to an unusually high mutation rate in animal mitochondria.

Molecular clocks have a finer time resolution than the fossil record and are a more reliable guide to the detailed structure of phylogenetic trees than are classical methods of tree construction, which are based on comparisons of the morphology and development of different species. For example, the precise relationship among the great-ape and human lineages was not settled until sufficient molecular-sequence data accumulated in the 1980s to produce the tree that was shown in Figure 4–75. And with huge amounts of DNA sequence now determined from a variety of mammals, much better estimates of our relationship to them are being obtained (**Figure 4–77**).

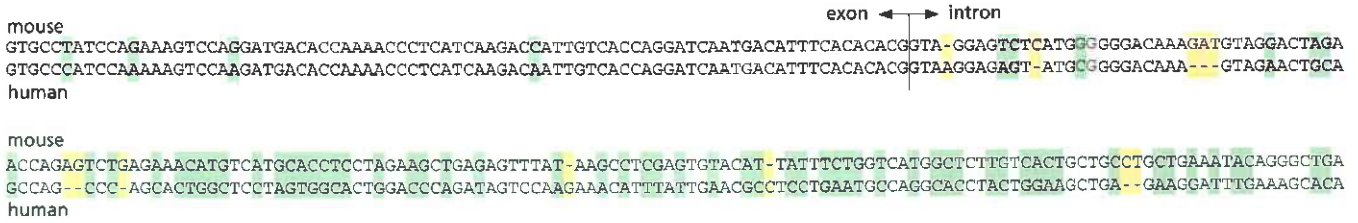
A Comparison of Human and Mouse Chromosomes Shows How The Structures of Genomes Diverge

As would be expected, the human and chimpanzee genomes are much more alike than are the human and mouse genomes. Although the size of the human and mouse genomes are roughly the same and they contain nearly identical sets of genes, there has been a much longer time period over which changes have had a chance to accumulate—approximately 80 million years versus 6 million years. In addition, as indicated in Figure 4–77, rodent lineages (represented by the rat and the mouse) have unusually fast molecular clocks. Hence, these lineages have diverged from the human lineage more rapidly than otherwise expected.

As indicated by the DNA sequence comparison in **Figure 4–78**, mutation has led to extensive sequence divergence between humans and mice at all sites that are not under selection—such as most nucleotide sequences in introns. In contrast, in human–chimpanzee comparisons, nearly all sequence positions are the same simply because not enough time has elapsed since the last common ancestor for large numbers of changes to have occurred.

In contrast to the situation for humans and chimpanzees, local gene order and overall chromosome organization have diverged greatly between humans

Figure 4–78 Comparison of a portion of the mouse and human leptin genes. Positions where the sequences differ by a single nucleotide substitution are boxed in *green*, and positions that differ by the addition or deletion of nucleotides are boxed in *yellow*. Note that the coding sequence of the exon is much more conserved than is the adjacent intron sequence.



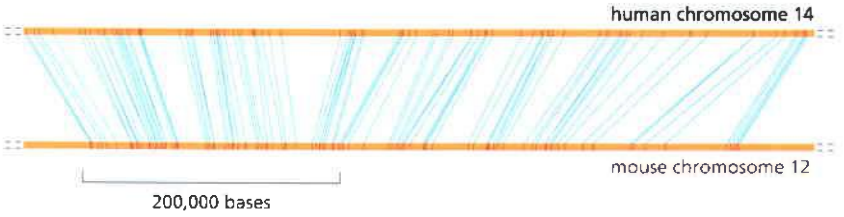


Figure 4–79 Comparison of a syntenic portion of mouse and human genomes. About 90 percent of the two genomes can be aligned in this way. Note that while there is an identical order of the matched index sequences (*red marks*), there has been a net loss of DNA in the mouse lineage that is interspersed throughout the entire region. This type of net loss is typical for all such regions, and it accounts for the fact that the mouse genome contains 14 percent less DNA than does the human genome. (Adapted from Mouse Sequencing Consortium, *Nature* 420:520–573, 2002. With permission from Macmillan Publishers Ltd.)

and mice. According to rough estimates, a total of about 180 break-and-rejoin events have occurred in the human and mouse lineages since these two species last shared a common ancestor. In the process, although the number of chromosomes is similar in the two species (23 per haploid genome in the human versus 20 in the mouse), their overall structures differ greatly. Nonetheless, even after the extensive genomic shuffling, there are many large blocks of DNA in which the gene order is the same in the human and the mouse. These stretches of conserved gene order in chromosomes are referred to as regions of *synteny.*

An unexpected conclusion from a detailed comparison of the complete mouse and human genome sequences, confirmed from subsequent comparisons between the genomes of other vertebrates, is that small blocks of sequences are being deleted from and added to genomes at a surprisingly rapid rate. Thus, if we assume that our common ancestor had a genome of human size (about 3 billion nucleotide pairs), mice would have lost a total of about 45 percent of that genome from accumulated deletions during the past 80 million years, while humans would have lost about 25 percent. However, substantial sequence gains from many small chromosome duplications and from the multiplication of transposons have compensated for these deletions. As a result, our genome size is unchanged from that of the last common ancestor for humans and mice, while the mouse genome is smaller by only 0.3 billion nucleotides.

Good evidence for the loss of DNA sequences in small blocks during evolution can be obtained from a detailed comparison of most regions of synteny in the human and mouse genomes. The comparative shrinkage of the mouse genome can be clearly seen from such comparisons, with the net loss of sequences scattered throughout the long stretches of DNA that are otherwise homologous (**Figure 4–79**).

DNA is added to genomes both by the spontaneous duplication of chromosomal segments that contain tens of thousands of nucleotide pairs (as will be discussed shortly), and by active transposition (most transposition events are duplicative, because the original copy of the transposon stays where it was when a copy inserts at the new site; for example, see Figure 5–74). Comparison of the DNA sequences derived from transposons in the human and the mouse therefore readily reveals some of the sequence additions (**Figure 4–80**).

For unknown reasons, all mammals have genome sizes of about 3 billion nucleotide pairs that contain nearly identical sets of genes, even though only on the order of 150 million nucleotide pairs appear to be under sequence-specific functional constraints.

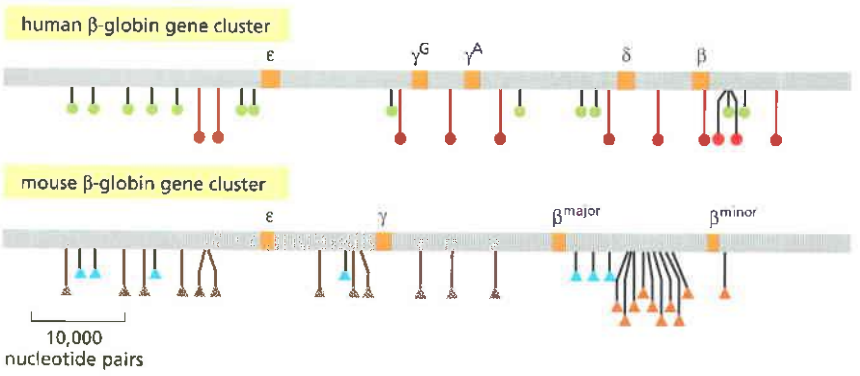


Figure 4–80 A comparison of the β-globin gene cluster in the human and mouse genomes, showing the location of transposable elements. This stretch of human genome contains five functional β-globin-like genes *(orange);* the comparable region from the mouse genome has only four. The positions of the human Alu sequence are indicated by green circles, and the human L1 sequences by red circles. The mouse genome contains different but related transposable elements: the positions of B1 elements (which are related to the human Alu sequences) are indicated by *blue triangles,* and the positions of the mouse L1 elements (which are related to the human L1 sequences) are indicated by *orange triangles.* The absence of transposable elements from the globin structural genes can be attributed to purifying selection, which would have eliminated any insertion that compromised gene function. (Courtesy of Ross Hardison and Webb Miller.)

The Size of a Vertebrate Genome Reflects the Relative Rates of DNA Addition and DNA Loss in a Lineage

Now that we know the complete sequence of a number of vertebrate genomes, we see that genome size can vary considerably, apparently without a drastic effect on the organism or its number of genes. Thus, while the mouse and dog genomes are both in the typical mammalian size range, the chicken has a genome that is only about one-third human size (one billion nucleotide pairs). A particularly notable example of an organism with a genome of anomalous size is the puffer fish, *Fugu rubripes* (**Figure 4–81**), which has a tiny genome for a vertebrate (0.4 billion nucleotide pairs compared to 1 billion or more for many other fish). The small size of the *Fugu* genome is largely due to the small size of its introns. Specifically, *Fugu* introns, as well as other noncoding segments of the *Fugu* genome, lack the repetitive DNA that makes up a large portion of the genomes of most well-studied vertebrates. Nevertheless, the positions of *Fugu* introns are nearly perfectly conserved relative to their positions in mammalian genomes (**Figure 4–82**).

While initially a mystery, we now have a simple explanation for such large differences in genome size between similar organisms: because all vertebrates experience a continuous process of DNA loss and DNA addition, the size of a genome merely depends on the balance between these opposing processes acting over millions of years. Suppose, for example, that in the lineage leading to *Fugu*, the rate of DNA addition happened to slow greatly. Over long periods of time, this would result in a major "cleansing" from this fish genome of those DNA sequences whose loss could be tolerated. In retrospect, the process of purifying selection in the *Fugu* lineage has partitioned those vertebrate DNA sequences most likely to be functional into only 400 million nucleotide pairs of DNA, providing a major resource for scientists.

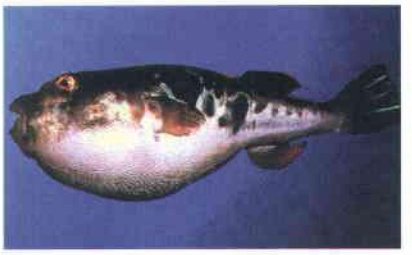


Figure 4–81 The puffer fish, *Fugu rubripes*. (Courtesy of Byrappa Venkatesh.)

We Can Reconstruct the Sequence of Some Ancient Genomes

The genomes of ancestral organisms can be inferred, but never directly observed: there are no ancient organisms alive today. Although a modern organism such as the horseshoe crab looks remarkably similar to fossil ancestors that lived 200 million years ago, there is every reason to believe that the horseshoe-crab genome has been changing during all that time at a rate similar to that occurring in other evolutionary lineages. Selection constraints must have maintained key functional properties of the horseshoe-crab genome to account for the morphological stability of the lineage. However, genome sequences reveal that the fraction of the genome subject to purifying selection is small; hence the genome of the modern horseshoe crab must differ greatly from that of its extinct ancestors, known to us only through the fossil record.

Is there any way around this problem? Can we ever hope to decipher large sections of the genome sequence of the extinct ancestors of organisms that are

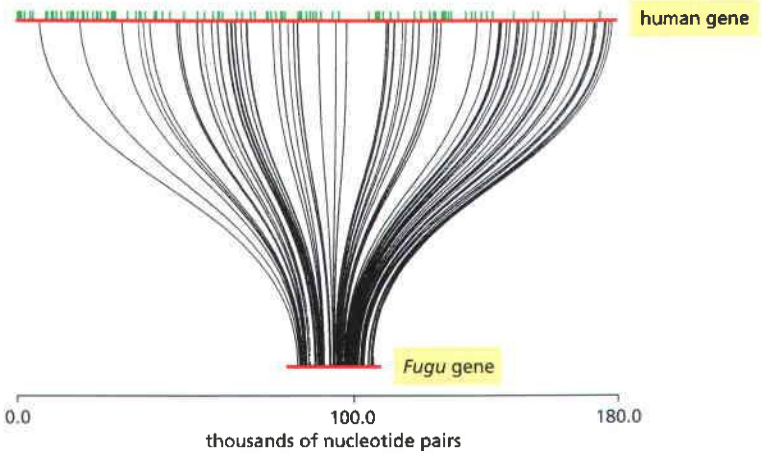


Figure 4–82 Comparison of the genomic sequences of the human and *Fugu* genes encoding the protein huntingtin. Both genes (indicated in *red*) contain 67 short exons that align in 1:1 correspondence to one another; these exons are connected by curved lines. The human gene is 7.5 times larger than the *Fugu* gene (180,000 versus 27,000 nucleotide pairs). The size difference is entirely due to larger introns in the human gene. The larger size of the human introns is due in part to the presence of retrotransposons, whose positions are represented by green vertical lines; the *Fugu* introns lack retrotransposons. In humans, mutation of the huntingtin gene causes Huntington's disease, an inherited neurodegenerative disorder. (Adapted from S. Baxendale et al., *Nat. Genet.* 10:67–76, 1995. With permission from Macmillan Publishers Ltd.)

alive today? For organisms that are as closely related as human and chimp, we saw that this may not be difficult. In that case, reference to the gorilla sequence can be be used to sort out which of the few differences between human and chimp DNA sequences was inherited from our common ancestor some 6 million years ago (see Figure 4–76). For an ancestor that has produced a large number of different organisms alive today, the DNA sequences of many species can be compared simultaneously to unscramble the ancestral sequence, allowing scientists to trace DNA sequences much farther back in time. For example, from the complete genome sequences of 20 modern mammals that will soon be obtained, it should be possible to decipher most of the genome sequence of the 100 million year-old Boreoeutherian mammal that gave rise to species as diverse as dog, mouse, rabbit, armadillo and human (see Figure 4–77).

Multispecies Sequence Comparisons Identify Important DNA Sequences of Unknown Function

The massive quantity of DNA sequence now in databases (more than a hundred billion nucleotide pairs) provides a rich resource that scientists can mine for many purposes. We have already discussed how this information can be used to unscramble the evolutionary pathways that have led to modern organisms. But sequence comparisons also provide many insights into how cells and organisms function. Perhaps the most remarkable discovery in this realm has been the observation that, although only about 1.5% of the human genome codes for proteins, about three times this amount (in total, 5% of the genome—see Table 4–1, p. 206) has been strongly conserved during mammalian evolution. This mass of conserved sequence is most clearly revealed when we align and compare DNA synteny blocks from many different species. In this way, so-called *multispecies conserved sequences* can be readily identified (**Figure 4–83**). Most of the noncoding conserved sequences discovered in this way turn out to be relatively short, containing between 50 and 200 nucleotide pairs. The strict conservation

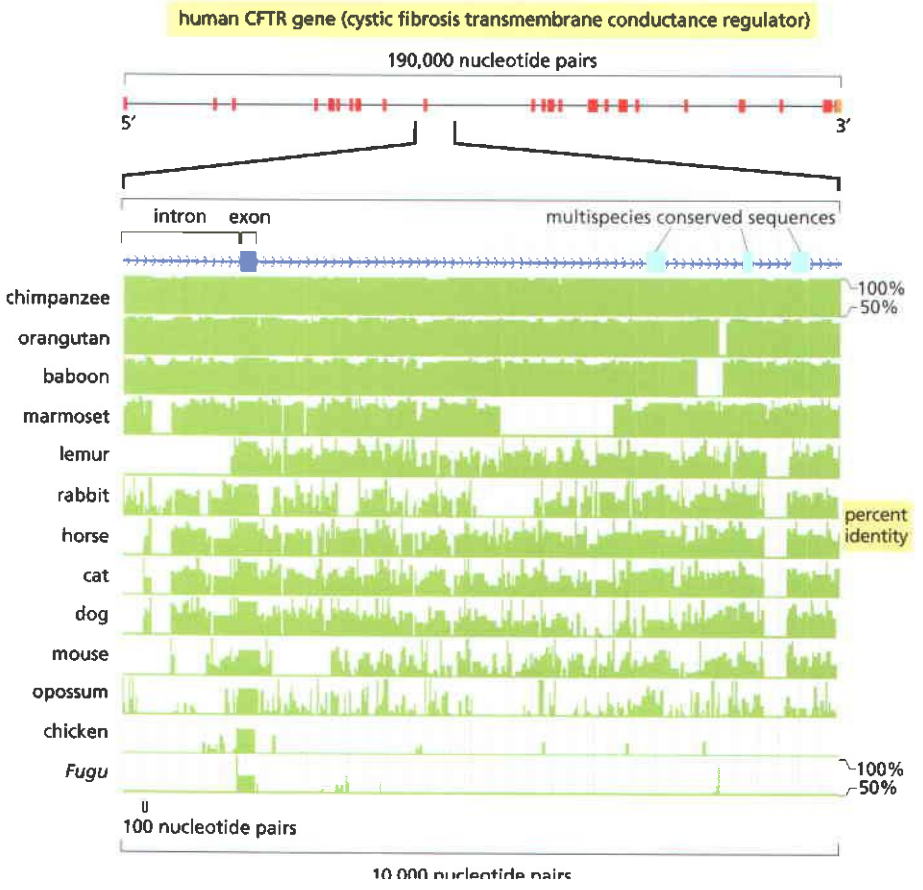


Figure 4–83 The detection of multispecies conserved sequences. In this example, genome sequences for each of the organisms shown have been compared with the indicated region of the human CFTR gene, scanning in 25 nucleotide blocks. For each organism, the percent identity across its syntenic sequences is plotted in *green*. In addition, a computational algorithm has been used to detect the sequences within this region that are most highly conserved when the sequences from all of the organisms are taken into account. Besides the exon, three other blocks of multispecies conserved sequences are shown. The function of most such sequences in the human genome is not known. (Courtesy of Eric D. Green.)

implies that they have important functions that have been maintained by purifying selection. The puzzle is to unravel what those functions are. Some of the conserved sequence that does not code for protein codes for untranslated RNA molecules that are known to have important functions, as we shall see in later chapters. Another fraction of the noncoding conserved DNA is certainly involved in regulating the transcription of adjacent genes, as discussed in Chapter 7. But we do not yet know how much of the conserved DNA can be accounted for in these ways, and the bulk of it is still a deep mystery. The solution to this mystery is bound to have profound consequences for medicine, and it reveals how much more we need to learn about the biology of vertebrate organisms.

How can cell biologists tackle this problem? The first step is to distinguish between the conserved regions that code for protein and those that do not, and then, among the latter, to focus on those that do not already have some other identified function, in coding for structural RNA molecules, for example. The next task is to discover what proteins or RNA molecules bind to these mysterious DNA sequences, how they are packaged into chromatin, and whether they ever serve as templates for RNA synthesis. Most of this task still lies before us, but a start has been made, and some remarkable insights have been obtained. One of the most intriguing concerns the evolutionary changes that have made us humans different from other animals—changes, that is, in sequences that have been conserved among our close relatives but have undergone sudden rapid change in the human sublineage.

Accelerated Changes in Previously Conserved Sequences Can Help Decipher Critical Steps in Human Evolution

As soon as both the human and the chimpanzee genome sequences became available, scientists began searching for DNA sequence changes that might account for the striking differences between us and them. With 3 billion nucleotide pairs to compare in the two species, this might seem an impossible task. But the job was made much easier by confining the search to 35,000 clearly defined multispecies conserved sequences (a total of about 5 million nucleotide pairs), representing parts of the genome that are most likely to be functionally important. Though these sequences are conserved strongly, they are not conserved perfectly, and when the version in one species is compared with that in another they are generally found to have drifted apart by a small amount corresponding simply to the time elapsed since the last common ancestor. In a small proportion of cases, however, one sees signs of a sudden evolutionary spurt. For example, some DNA sequences that have been highly conserved in other mammalian species are found to have changed exceptionally fast during the six million years of human evolution since we diverged from the chimpanzees. Such *human accelerated regions* (HARs) are thought to reflect functions that have been especially important in making us different in some useful way.

About 50 such sites were identified in one study, one-fourth of which were located near genes associated with neural development. The sequence exhibiting the most rapid change (18 changes between human and chimp, compared to only two changes between chimp and chicken) was examined further and found to encode a 118-nucleotide noncoding RNA molecule that is produced in the human cerebral cortex at a critical time during brain development (**Figure 4–84**). Although the function of this HAR1F RNA is not yet known, this exciting finding is stimulating further studies that will hopefully shed light on crucial features of the human brain.

Gene Duplication Provides an Important Source of Genetic Novelty During Evolution

Evolution depends on the creation of new genes, as well as on the modification of those that already exist. How does this occur? When we compare organisms that seem very different—a primate with a rodent, for example, or a mouse with

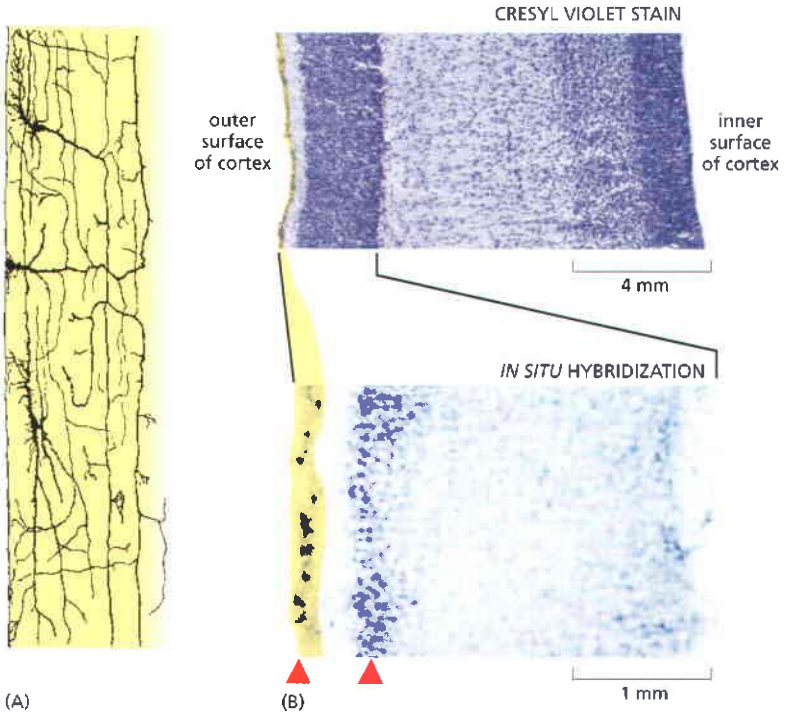


Figure 4–84 Initial characterization of a new gene detected as a previously conserved DNA sequence that evolved rapidly in humans. (A) Drawing by Ramon y Cajal of the outer surface of the human neocortex, highlighting the Cajal–Retzius neurons. (B) Tissue slices from an embryonic human brain showing part of the cortex, with the region containing the Cajal–Retzius neurons highlighted in *yellow*. Upper photograph: cresyl violet stain. Lower photograph: *in situ* hybridization. The *red arrows* indicate the cells that produce HAR1F RNA as detected by *in situ* hybridization *(blue)*. HAR1F is a novel noncoding RNA that has evolved rapidly in the human lineage leading from the great apes. The Cajal–Retzius neurons make this RNA at the time when the neocortex is developing. The results are intriguing, because a large neocortex is special to humans; for the behavior of cells in forming this cortex, see Figure 22–99. (Adapted from K.S. Pollard et al., *Nature* 443:167–172, 2006. With permission from Macmillan Publishers Ltd.)

a fish—we rarely encounter genes in the one species that have no homolog in the other. Genes without homologous counterparts are relatively scarce even when we compare such divergent organisms as a mammal and a worm. On the other hand, we frequently find gene families that have different numbers of members in different species. To create such families, genes have been repeatedly duplicated, and the copies have then diverged to take on new functions that often vary from one species to another.

The genes encoding nuclear hormone receptors in humans, a nematode worm, and a fruit fly illustrate this point (**Figure 4–85**). Many of the subtypes of these nuclear receptors (also called intracellular receptors) have close homologs in all three organisms that are more similar to each other than they are to other family subtypes present in the same species. Therefore, much of the functional divergence of this large gene family must have preceded the divergence of these three evolutionary lineages. Subsequently, one major branch of the gene family underwent an enormous expansion in the worm lineage only. Similar, but smaller, lineage-specific expansions of particular subtypes are evident throughout the gene family tree.

Gene duplication occurs at high rates in all evolutionary lineages, contributing to the vigorous process of DNA addition discussed previously. In a detailed study of spontaneous duplications in yeast, duplications of 50,000 to 250,000 nucleotide pairs were commonly observed, most of which were tandemly repeated. These appeared to result from DNA replication errors that led to the inexact repair of double-strand chromosome breaks. A comparison of the human and chimpanzee genomes reveals that, since the time that these two organisms diverged, segmental duplications have added about 5 million nucleotide pairs to each genome every million years, with an average duplication size being about 50,000 nucleotide pairs (however, there are duplications five times larger, as in yeast). In fact, if one counts nucleotides, duplication events have created more differences between our two species than have single nucleotide substitutions.

Duplicated Genes Diverge

A major question in genome evolution concerns the fate of newly duplicated genes. In most cases, there is presumed to be little or no selection—at least initially—to maintain the duplicated state since either copy can provide an equiv-

alent function. Hence, many duplication events are likely to be followed by loss-of-function mutations in one or the other gene. This cycle would functionally restore the one-gene state that preceded the duplication. Indeed, there are many examples in contemporary genomes where one copy of a duplicated gene can be seen to have become irreversibly inactivated by multiple mutations. Over time, the sequence similarity between such a **pseudogene** and the functional gene whose duplication produced it would be expected to be eroded by the accumulation of many mutations in the pseudogene—the homologous relationship eventually becoming undetectable.

An alternative fate for gene duplications is for both copies to remain functional, while diverging in their sequence and pattern of expression, thus taking on different roles. This process of "duplication and divergence" almost certainly explains the presence of large families of genes with related functions in biologically complex organisms, and it is thought to play a critical role in the evolution of increased biological complexity. An examination of many different eucaryotic genomes suggests that the probability that any particular gene will undergo a duplication event that spreads to most or all individuals in a species is approximately 1% every million years.

Whole-genome duplications offer particularly dramatic examples of the duplication–divergence cycle. A whole-genome duplication can occur quite simply: all that is required is one round of genome replication in a germline cell lineage without a corresponding cell division. Initially, the chromosome number simply doubles. Such abrupt increases in the ploidy of an organism are common, particularly in fungi and plants. After a whole-genome duplication, all genes exist as duplicate copies. However, unless the duplication event occurred so recently that there has been little time for subsequent alterations in genome structure, the results of a series of segmental duplications—occurring at different times—are very hard to distinguish from the end product of a whole-genome duplication. In mammals, for example, the role of whole-genome duplications versus a series of piecemeal duplications of DNA segments is quite uncertain. Nevertheless, it is clear that a great deal of gene duplication has ocurred in the distant past.

Analysis of the genome of the zebrafish, in which either a whole-genome duplication or a series of more local duplications occurred hundreds of millions of years ago, has cast some light on the process of gene duplication and divergence. Although many duplicates of zebrafish genes appear to have been lost by mutation, a significant fraction—perhaps as many as 30–50%—have diverged functionally while both copies have remained active. In many cases, the most obvious functional difference between the duplicated genes is that they are expressed in different tissues or at different stages of development (see Figure 22–46). One attractive theory to explain such an end result imagines that different, mildly deleterious mutations occur quickly in both copies of a duplicated gene set. For example, one copy might lose expression in a particular tissue as a result of a regulatory mutation, while the other copy loses expression in a second tissue. Following such an occurrence, both gene copies would be required to provide the full range of functions that were once supplied by a single gene; hence, both copies would now be protected from loss through inactivating mutations. Over a longer period, each copy could then undergo further changes through which it could acquire new, specialized features.

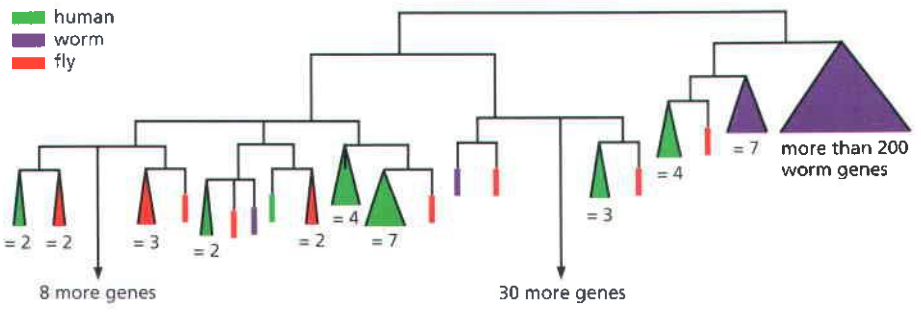


Figure 4–85 A phylogenetic tree based on the inferred protein sequences for all nuclear hormone receptors encoded in the genomes of human (*H. sapiens*), a nematode worm (*C. elegans*), and a fruit fly (*D. melanogaster*). Triangles represent protein subfamilies that have expanded within individual evolutionary lineages; the width of these triangles indicates the number of genes encoding members of these subfamilies. Colored vertical bars represent a single gene. There is no simple pattern to the historical duplications and divergences that have created the gene families encoding nuclear receptors in the three contemporary organisms. The family of nuclear hormone receptors is described in Figure 15–14. These proteins function in cell signaling and gene regulation. (Adapted from International Human Genome Sequencing Consortium, *Nature* 409:860–921, 2001. With permission from Macmillan Publishers Ltd.)

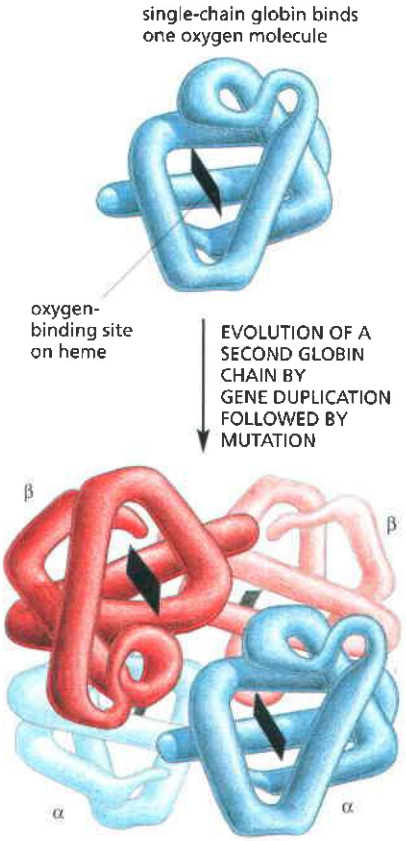


Figure 4–86 A comparison of the structure of one-chain and four-chain globins. The four-chain globin shown is hemoglobin, which is a complex of two α-globin and two β-globin chains. The one-chain globin in some primitive vertebrates forms a dimer that dissociates when it binds oxygen, representing an intermediate in the evolution of the four-chain globin.

The Evolution of the Globin Gene Family Shows How DNA Duplications Contribute to the Evolution of Organisms

The globin gene family provides an especially good example of how DNA duplication generates new proteins, because its evolutionary history has been worked out particularly well. The unmistakable similarities in amino acid sequence and structure among the present-day globins indicate that they all must derive from a common ancestral gene, even though some are now encoded by widely separated genes in the mammalian genome.

We can reconstruct some of the past events that produced the various types of oxygen-carrying hemoglobin molecules by considering the different forms of the protein in organisms at different positions on the phylogenetic tree of life. A molecule like hemoglobin was necessary to allow multicellular animals to grow to a large size, since large animals could no longer rely on the simple diffusion of oxygen through the body surface to oxygenate their tissues adequately. Consequently, hemoglobin-like molecules are found in all vertebrates and in many invertebrates. The most primitive oxygen-carrying molecule in animals is a globin polypeptide chain of about 150 amino acids, which is found in many marine worms, insects, and primitive fish. The hemoglobin molecule in more complex vertebrates, however, is composed of two kinds of globin chains. It appears that about 500 million years ago, during the continuing evolution of fish, a series of gene mutations and duplications occurred. These events established two slightly different globin genes, coding for the α- and β-globin chains, in the genome of each individual. In modern vertebrates, each hemoglobin molecule is a complex of two α chains and two β chains (**Figure 4–86**). The four oxygen-binding sites in the $\alpha_2\beta_2$ molecule interact, allowing a cooperative allosteric change in the molecule as it binds and releases oxygen, which enables hemoglobin to take up and release oxygen more efficiently than the single-chain version.

Still later, during the evolution of mammals, the β-chain gene apparently underwent duplication and mutation to give rise to a second β-like chain that is synthesized specifically in the fetus. The resulting hemoglobin molecule has a higher affinity for oxygen than adult hemoglobin and thus helps in the transfer of oxygen from the mother to the fetus. The gene for the new β-like chain subsequently duplicated and mutated again to produce two new genes, ε and γ, the ε chain being produced earlier in development (to form $\alpha_2\varepsilon_2$) than the fetal γ chain, which forms $\alpha_2\gamma_2$. A duplication of the adult β-chain gene occurred still later, during primate evolution, to give rise to a δ-globin gene and thus to a minor form of hemoglobin ($\alpha_2\delta_2$) that is found only in adult primates (**Figure 4–87**).

Each of these duplicated genes has been modified by point mutations that affect the properties of the final hemoglobin molecule, as well as by changes in regulatory regions that determine the timing and level of expression of the gene. As a result, each globin is made in different amounts at different times of human development (see Figure 7–64B).

The end result of the gene duplication processes that have given rise to the diversity of globin chains is seen clearly in the human genes that arose from

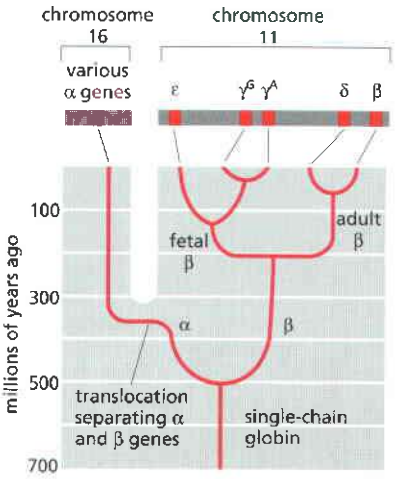


Figure 4–87 An evolutionary scheme for the globin chains that carry oxygen in the blood of animals. The scheme emphasizes the β-like globin gene family. A relatively recent gene duplication of the γ-chain gene produced γ^G and γ^A, which are fetal β-like chains of identical function. The location of the globin genes in the human genome is shown at the top of the figure (see also Figure 7–64).

the original β gene, which are arranged as a series of homologous DNA sequences located within 50,000 nucleotide pairs of one another. A similar cluster of α-globin genes is located on a separate human chromosome. Because the α- and β-globin gene clusters are on separate chromosomes in birds and mammals but are together in the frog *Xenopus*, it is believed that a chromosome translocation event separated the two gene clusters about 300 million years ago (see Figure 4–87).

There are several duplicated globin DNA sequences in the α- and β-globin gene clusters that are not functional genes but pseudogenes. These have a close sequence similarity to the functional genes but have been disabled by mutations that prevent their expression. The existence of such pseudogenes makes it clear that, as expected, not every DNA duplication leads to a new functional gene. We also know that nonfunctional DNA sequences are not rapidly discarded, as indicated by the large excess of noncoding DNA that is found in mammalian genomes.

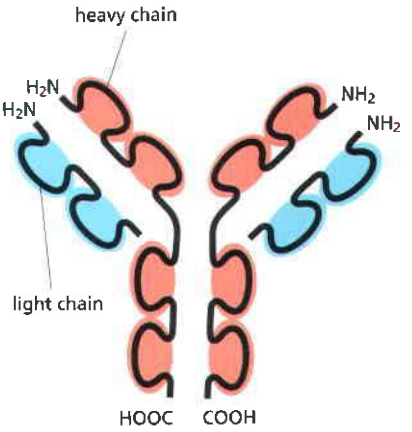


Figure 4–88 Schematic view of an antibody (immunoglobulin) molecule. This molecule is a complex of two identical heavy chains and two identical light chains. Each heavy chain contains four similar, covalently linked domains, while each light chain contains two such domains. Each domain is encoded by a separate exon, and all of the exons are thought to have evolved by the serial duplication of a single ancestral exon.

Genes Encoding New Proteins Can Be Created by the Recombination of Exons

The role of DNA duplication in evolution is not confined to the expansion of gene families. It can also act on a smaller scale to create single genes by stringing together short duplicated segments of DNA. The proteins encoded by genes generated in this way can be recognized by the presence of repeating similar protein domains, which are covalently linked to one another in series. The immunoglobulins (**Figure 4–88**) and albumins, for example, as well as most fibrous proteins (such as collagens) are encoded by genes that have evolved by repeated duplications of a primordial DNA sequence.

In genes that have evolved in this way, as well as in many other genes, each separate exon often encodes an individual protein folding unit, or domain. It is believed that the organization of DNA coding sequences as a series of such exons separated by long introns has greatly facilitated the evolution of new proteins. The duplications necessary to form a single gene coding for a protein with repeating domains, for example, can often occur by breaking and rejoining the DNA anywhere in the long introns on either side of an exon; without introns there would be only a few sites in the original gene at which a recombinational exchange between DNA molecules could duplicate the domain. By enabling the duplication to occur by recombination at many potential sites rather than just a few, introns increase the probability of a favorable duplication event.

More generally, we know from genome sequences that the various parts of genes—both their individual exons and their regulatory elements—have served as modular elements that have been duplicated and moved about the genome to create the great diversity of living things. Thus, for example, many present-day proteins are formed as a patchwork of domains from different origins, reflecting their long evolutionary history (see Figure 3–19).

Neutral Mutations Often Spread to Become Fixed in a Population, with a Probability that Depends on Population Size

In comparisons between two species that have diverged from one another by millions of years, it makes little difference which individuals from each species are compared. For example, typical human and chimpanzee DNA sequences differ from one another by about 1%. In contrast, when the same region of the genome is sampled from two different humans, the differences are typically less than 0.1%. For more distantly related organisms, the inter-species differences overshadow intra-species variation even more dramatically. However, each "fixed difference" between the human and the chimpanzee (in other words, each difference that is now characteristic of all or nearly all individuals of each species) started out as a new mutation in a single individual. If the size of the

interbreeding population in which the mutation occurred is N, the initial allele frequency of a new mutation would be $1/(2N)$ for a diploid organism. How does such a rare mutation become fixed in the population, and hence become a characteristic of the species rather than of a particular individual genome?

The answer to this question depends on the functional consequences of the mutation. If the mutation has a significantly deleterious effect, it will simply be eliminated by purifying selection and will not become fixed. (In the most extreme case, the individual carrying the mutation will die without producing progeny.) Conversely, the rare mutations that confer a major reproductive advantage on individuals who inherit them can spread rapidly in the population. Because humans reproduce sexually and genetic recombination occurs each time a gamete is formed (discussed in Chapter 5), the genome of each individual who has inherited the mutation will be a unique recombinational mosaic of segments inherited from a large number of ancestors. The selected mutation along with a modest amount of neighboring sequence—ultimately inherited from the individual in which the mutation occurred—will simply be one piece of this huge mosaic.

The great majority of mutations that are not harmful are not beneficial either. These selectively neutral mutations can also spread and become fixed in a population, and they make a large contribution to the evolutionary change in genomes. Their spread is not as rapid as the spread of the rare strongly advantageous mutations. The process by which such neutral genetic variation is passed down through an idealized interbreeding population can be described mathematically by equations that are surprisingly simple. The idealized model that has proven most useful for analyzing human genetic variation assumes a constant population size and random mating, as well as selective neutrality for the mutations. While neither of the first two assumptions is a good description of human population history, they nonetheless provide a useful starting point for analyzing intra-species variation.

When a new neutral mutation occurs in a constant population of size N that is undergoing random mating, the probability that it will ultimately become fixed is approximately $1/(2N)$. For those mutations that do become fixed, the average time to fixation is approximately $4N$ generations. A detailed analysis of data on human genetic variation suggests an ancestral population size of approximately 10,000 during the period when the current pattern of genetic variation was largely established. With a population that has reached this size, the probability that a new, selectively neutral mutation would become fixed is small (5×10^{-5}), while the average time to fixation would be on the order of 800,000 years (assuming a 20-year generation time). Thus, while we know that the human population has grown enormously since the development of agriculture approximately 15,000 years ago, most of the present-day set of common human genetic variants reflects the mixture of variants that was already present long before this time, when the human population was still small enough to allow their widespread dissemination.

A Great Deal Can Be Learned from Analyses of the Variation Among Humans

Even though most of the variation among modern humans originates from variation present in a comparatively tiny group of ancestors, the number of variations encountered is very large. One important source of variation, which was missed for many years, is the presence of many duplications and deletions of large blocks of DNA. According to one estimate, when any individual human is compared with the standard reference genome in the database, one should expect to find roughly 100 differences involving long sequence blocks. Some of these "copy number variations" will be very common (**Figure 4–89**), while others will be present in only a minority of humans (**Figure 4–90**). From an initial sampling, nearly half will contain known genes. In retrospect this type of variation is not surprising, given the extensive history of DNA addition and DNA loss in vertebrate genomes (for example, see Figure 4–79).

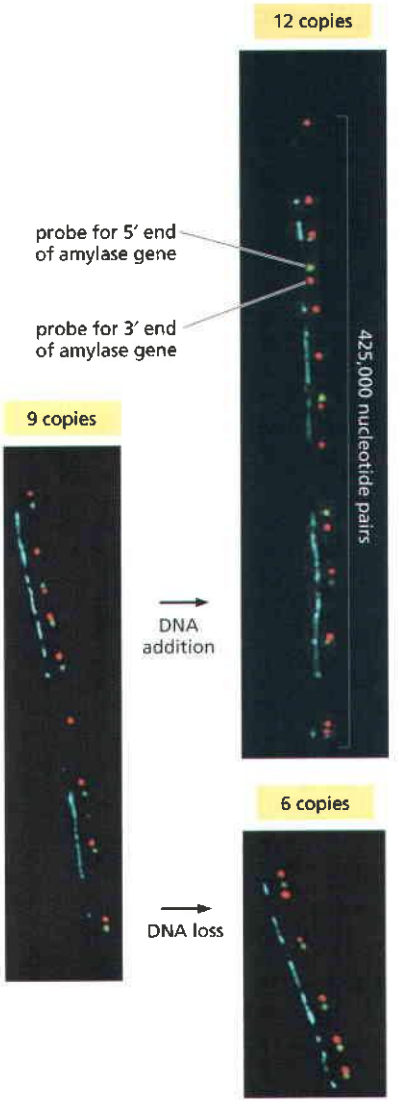


Figure 4–89 Visualization of a frequent type of variation among humans. About half of the humans tested had nine copies of the amylase gene (*left*), which produces an important enzyme that digests starch. In other humans, there has been either DNA loss or DNA addition to produce an altered chromosome, resulting from the deletion (loss) or the duplication (addition) of a part of this region. To obtain these images, stretched chromatin fibers have been hybridized with differently colored probes to the two ends of the amylase gene, as indicated. The *blue* lines mark the general paths of the chromatin. They have been determined by a second stain and displaced to one side for clarity. (Adapted from A.J. Iafrate et al., *Nat. Genet.* 36:949–951, 2004. With permission from Macmillan Publishers Ltd.)

The intra-species variations that have been most extensively characterized are **single-nucleotide polymorphisms** (**SNPs**). These are simply points in the genome sequence where one large fraction of the human population has one nucleotide, while another substantial fraction has another. Two human genomes sampled from the modern world population at random will differ at approximately 2.5×10^6 such sites (1 per 1300 nucleotide pairs). As will be described in the overview of genetics in Chapter 8, mapped sites in the human genome that are **polymorphic**—meaning that there is a reasonable probability (generally more than 1%) that the genomes of two individuals will differ at that site—are extremely useful for genetic analyses, in which one attempts to associate specific traits (phenotypes) with specific DNA sequences for medical or scientific purposes (see p. 560).

Against the background of ordinary SNPs inherited from our prehistoric ancestors, certain sequences with exceptionally high mutation rates stand out. A dramatic example is provided by CA repeats, which are ubiquitous in the human genome and in the genomes of other eucaryotes. Sequences with the motif $(CA)_n$ are replicated with relatively low fidelity because of a slippage that occurs between the template and the newly synthesized strands during DNA replication; hence, the precise value of n can vary over a considerable range from one genome to the next. These repeats make ideal DNA-based genetic markers, since most humans are heterozygous—carrying two values of n at any particular CA repeat, having inherited one repeat length (n) from their mother and a different repeat length from their father. While the value of n changes sufficiently rarely that most parent–child transmissions propagate CA repeats faithfully, the changes are sufficiently frequent to maintain high levels of heterozygosity in the human population. These and some other simple repeats that display exceptionally high variability therefore provide the basis for identifying individuals by DNA analysis in crime investigations, paternity suits, and other forensic applications (see Figure 8–47).

While most of the SNPs and copy number variations in the human genome sequence are thought to have no effect on phenotype, a subset of them must be

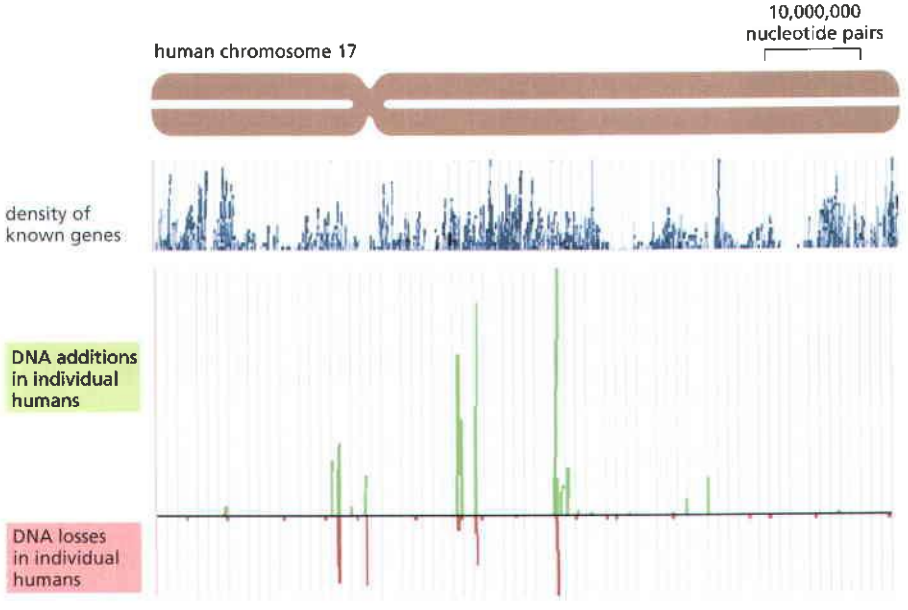


Figure 4–90 Detection of copy number variants on human chromosome 17. When 100 individuals were tested by a DNA microarray analysis that detects the copy number of DNA sequences throughout the entire length of this chromosome, the indicated distributions of DNA additions (*green bars*) and DNA losses (*red bars*) were observed compared with an arbitrary human reference sequence. The shortest *red* and *green bars* represent a single occurrence among the 200 chromosomes examined, whereas the longer bars indicate that the addition or loss was correspondingly more frequent. The results show preferred regions where the variations occur, and these tend to be in or near regions that already contain blocks of segmental duplications. Many of the changes include known genes. (Adapted from J.L. Freeman et al., *Genome Res.* 16:949–961, 2006. With permission from Cold Spring Harbor Laboratory Press.)

responsible for nearly all the heritable aspects of human individuality. We know that even a single nucleotide change that alters one amino acid in a protein can cause a serious disease, as for example in sickle cell anemia, which is caused by such a mutation in hemoglobin. <TTTT> We also know that gene dosage—a doubling or halving of the copy number of some genes—can have a profound effect on human development by altering the level of gene product. There is therefore every reason to suppose that some of the many differences between any two human beings will have substantial effects on human health, physiology, and behavior, whether they be SNPs or copy number variations. The major challenge in human genetics is to learn to recognize those relatively few variations that are functionally important against a large background of neutral variation in the genomes of different humans.

Summary

Comparisons of the nucleotide sequences of present-day genomes have revolutionized our understanding of gene and genome evolution. Because of the extremely high fidelity of DNA replication and DNA repair processes, random errors in maintaining the nucleotide sequences in genomes occur so rarely that only about one nucleotide in 1000 is altered every million years in any particular line of descent. Not surprisingly, therefore, a comparison of human and chimpanzee chromosomes—which are separated by about 6 million years of evolution—reveals very few changes. Not only are our genes essentially the same, but their order on each chromosome is almost identical. Although a substantial number of segmental duplications and segmental deletions have occurred in the past 6 million years, even the positions of the transposable elements that make up a major portion of our noncoding DNA are mostly unchanged.

When one compares the genomes of two more distantly related organisms—such as a human and a mouse, separated by about 80 million years—one finds many more changes. Now the effects of natural selection can be clearly seen: through purifying selection, essential nucleotide sequences—both in regulatory regions and in coding sequences (exon sequences)—have been highly conserved. In contrast, nonessential sequences (for example, much of the DNA in introns) have been altered to such an extent that an accurate alignment according to ancestry is frequently not possible.

Because of purifying selection, the comparison of the genome sequences of multiple related species is an especially powerful way to find DNA sequences with important functions. Although about 5% of the human genome has been conserved as a result of purifying selection, the function of the majority of this DNA (tens of thousands of multispecies conserved sequences) remains mysterious. Future experiments characterizing their functions should teach us a great deal about vertebrate biology.

Other sequence comparisons show that a great deal of the genetic complexity of present-day organisms is due to the expansion of ancient gene families. DNA duplication followed by sequence divergence has clearly been a major source of genetic novelty during evolution. The genomes of any two humans will differ from each other both because of nucleotide substitutions (single nucleotide polymorphisms, or SNPs) and because of inherited DNA gains and DNA losses that cause copy number variants. Understanding these differences will improve both medicine and our understanding of human biology.

PROBLEMS

Which statements are true? Explain why or why not.

4–1 Human females have 23 different chromosomes, whereas human males have 24.

4–2 In a comparison between the DNAs of related organisms such as humans and mice, identifying the conserved DNA sequences facilitates the search for functionally important regions.

4–3 The four core histones are relatively small proteins with a very high proportion of positively charged amino acids; the positive charge helps the histones bind tightly to DNA, regardless of its nucleotide sequence.

4–4 Nucleosomes bind DNA so tightly that they cannot move from the positions where they are first assembled.

4–5 Gene duplication and divergence is thought to have played a critical role in the evolution of increased biological complexity.

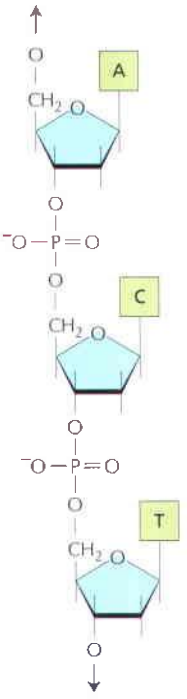


Figure Q4–1 Three nucleotides from the interior of a single strand of DNA (Problem 4–7). *Arrows* at the ends of the DNA strand indicate that the structure continues in both directions.

Discuss the following problems.

4–6 DNA isolated from the bacterial virus M13 contains 25% A, 33% T, 22% C, and 20% G. Do these results strike you as peculiar? Why or why not? How might you explain these values?

4–7 A segment of DNA from the interior of a single strand is shown in **Figure Q4–1**. What is the polarity of this DNA from top to bottom?

4–8 Human DNA contains 20% C on a molar basis. What are the mole percents of A, G, and T?

4–9 Chromosome 3 in orangutans differs from chromosome 3 in humans by two inversion events (**Figure Q4–2**). Draw the intermediate chromosome that resulted from the first inversion and explicitly indicate the segments included in each inversion.

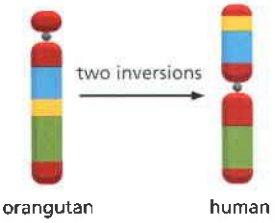


Figure Q4–2 Chromosome 3 in orangutans and humans (Problem 4–9). Differently *colored blocks* indicate segments of the chromosomes that were derived by previous fusions.

4–10 Assuming that the 30-nm chromatin fiber contains about 20 nucleosomes (200 bp/nucleosome) per 50 nm of length, calculate the degree of compaction of DNA associated with this level of chromatin structure. What fraction of the 10,000-fold condensation that occurs at mitosis does this level of DNA packing represent?

4–11 In contrast to histone acetylation, which always correlates with gene activation, histone methylation can lead to either transcriptional activation or repression. How do you suppose that the same modification—methylation—can mediate different biological outcomes?

4–12 Why is a chromosome with two centromeres (a dicentric chromosome) unstable? Would a back-up centromere not be a good thing for a chromosome, giving it two chances to form a kinetochore and attach to microtubules during mitosis? Would that not help to ensure that the chromosome did not get left behind at mitosis?

4–13 HP1 proteins, a family of proteins found in heterochromatin, are implicated in gene silencing and chromatin structure. The three proteins in humans—HP1α, HP1β, and HP1γ—share a highly conserved chromodomain, which is thought to direct chromatin localization. To determine whether these proteins could bind to the histone H3 N-terminus, you have covalently attached to separate beads various versions of the H3 N-terminal peptide—unmodified,

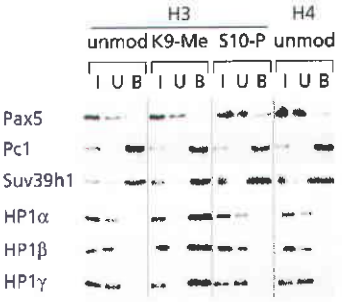


Figure Q4–3 Pull-down assays to determine binding specificity of HP1 proteins (Problem 4–13). Each protein at the left was detected by immunoblotting using a specific antibody after separation by SDS-polyacrylamide gel electrophoresis. For each histone N-terminal peptide the total input protein (I), the unbound protein (U), and the bound protein (B) are indicated. (Adapted from M. Lachner et al., *Nature* 410:116–120, 2001. With permission from Macmillan Publishers Ltd.)

Lys-9-dimethylated (K9-Me), and Ser-10-phosphorylated (S10-P)—along with an unmodified tail from histone H4. This arrangement allows you to incubate the beads with various proteins, wash away unbound proteins, and then elute bound proteins for assay by Western blotting. The results of your 'pull-down' assay for the HP1 proteins are shown in **Figure Q4–3**, along with the results from several control proteins, including Pax5, a gene regulatory protein, polycomb protein Pc1, which is known to bind to histones, and Suv39h1, a histone methyltransferase.

Based on these results, which of the proteins tested bind to the unmodified tails of histones? Do any of the HP1 proteins and control proteins selectively bind to the modified histone N-terminal peptides? What histone modification would you predict would be found in heterochromatin?

4–14 Mobile pieces of DNA—transposable elements—that insert themselves into chromosomes and accumulate during evolution make up more than 40% of the human genome. Transposable elements of four types—long interspersed elements (LINEs), short interspersed elements (SINEs), LTR retrotransposons, and DNA transposons—are inserted more or less randomly throughout the human genome. These elements are conspicuously rare at the four homeobox gene clusters, *HoxA*, *HoxB*, *HoxC*, and *HoxD*, as illustrated for *HoxD* in **Figure Q4–4**, along with an equivalent region of chromosome 22, which lacks a *Hox* cluster. Each *Hox* cluster is about 100 kb in length and contains 9 to 11 genes, whose differential expression along the anteroposterior axis of the developing embryo establishes the basic body plan for humans (and for other animals). Why do you suppose that transposable elements are so rare in the *Hox* clusters?



Figure Q4–4 Transposable elements and genes in 1 Mb regions of chromosomes 2 and 22 (Problem 4–14). Lines that project *upward* indicate exons of known genes. Lines that project *downward* indicate transposable elements; they are so numerous (constituting more than 40% of the human genome) that they merge into nearly a solid block outside the *Hox* clusters. (Adapted from E. Lander et al., *Nature* 409:860–921, 2001. With permission from Macmillan Publishers Ltd.)

REFERENCES

General

Hartwell L, Hood L, Goldberg ML et al (2006) Genetics: from Genes to Genomes, 3rd ed. Boston: McGraw Hill.

Olson MV (2002) The Human Genome Project: a player's perspective. *J Mol Biol* 319:931–942.

Strachan T & Read AP (2004) Human Molecular Genetics. New York: Garland Science.

Wolffe A (1999) Chromatin: Structure and Function, 3rd ed. New York: Academic Press.

The Structure and Function of DNA

Avery OT, MacLeod CM & McCarty M (1944) Studies on the chemical nature of the substance inducing transformation of *pneumococcal* types. *J Exp Med* 79:137–158.

Meselson M & Stahl FW (1958) The replication of DNA in *E. coli*. *Proc Natl Acad Sci USA* 44:671–682.

Watson JD & Crick FHC (1953) Molecular structure of nucleic acids. A structure for deoxyribose nucleic acids. *Nature* 171:737–738.

Chromosomal DNA and Its Packaging in the Chromatin Fiber

Jin J, Cai Y, Li B et al (2005) In and out: histone variant exchange in chromatin. *Trends Biochem Sci* 30:680–687.

Kornberg RD & Lorch Y (1999) Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. *Cell* 98:285–294.

Li G, Levitus M, Bustamante C & Widom J (2005) Rapid spontaneous accessibility of nucleosomal DNA. *Nature Struct Mol Biol* 12:46–53.

Lorch Y, Maier-Davis B & Kornberg RD (2006) Chromatin remodeling by nucleosome disassembly *in vitro*. *Proc Natl Acad Sci USA* 103:3090–3093.

Luger K, Mader AW, Richmond RK et al (1997) Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* 389:251–260.

Luger K & Richmond TJ (1998) The histone tails of the nucleosome. *Curr Opin Genet Dev* 8:140–146.

Malik H S & Henikoff S (2003) Phylogenomics of the nucleosome *Nature Struct Biol* 10:882-891.

Ried T, Schrock E, Ning Y & Wienberg J (1998) Chromosome painting: a useful art. *Hum Mol Genet* 7:1619–1626.

Robinson PJ & Rhodes R (2006) Structure of the 30 nm chromatin fibre: A key role for the linker histone. *Curr Opin Struct Biol* 16:1–8.

Saha A, Wittmeyer J & Cairns BR (2006) Chromatin remodeling: the industrial revolution of DNA around histones. *Nature Rev Mol Cell Biol* 7:437–446.

Woodcock CL (2006) Chromatin architecture. *Curr Opin Struct Biol* 16:213–220.

The Regulation of Chromatin Structure

Egger G, Liang, G, Aparicio A & Jones PA (2004) Epigenetics in human disease and prospects for epigenetic therapy. *Nature* 429:457–463.

Henikoff S (1990) Position-effect variegation after 60 years. *Trends Genet* 6:422–426.

Henikoff S & Ahmad K (2005) Assembly of variant histones into chromatiin. *Annu Rev Cell Dev Biol* 21:133–153.

Gaszner M & Felsenfeld G (2006) Insulators: exploiting transcriptional and epigenetic mechanisms. *Nature Rev Genet* 7:703–713.

Hake SB & Allis CD (2006) Histone H3 variants and their potential role in indexing mammalian genomes: the "H3 barcode hypothesis." *Proc Natl Acad Sci USA* 103:6428–6435.

Jenuwein T (2006) The epigenetic magic of histone lysine methylation. *FEBS J* 273:3121–3135.

Martin C & Zhang Y (2005) The diverse functions of histone lysine methylation. *Nature Rev Mol Cell Biol* 6:838–849.

Mellone B, Erhardt S & Karpen GH (2006) The ABCs of centromeres. *Nature Cell Biol* 8:427–429.

Peterson CL & Laniel MA (2004) Histones and histone modifications. *Curr Biol* 14:R546–R551.

Ruthenburg AJ, Allis CD & Wysocka J (2007) Methylation of lysine 4 on histone H3: intricacy of writing and reading a single epigenetic mark. *Mol Cell* 25:15–30.

Shahbazian MD & Grunstein M (2007) Functions of site-specific histone acetylation and deacetylation. *Annu Rev Biochem* 76:75–100.

The Global Structure of Chromosomes

Akhtar A & Gasser SM (2007) The nuclear envelope and transcriptional control. *Nature Rev Genet* 8:507–517.

Callan HG (1982) Lampbrush chromosomes. *Proc Roy Soc Lond Ser B* 21:417–448.

Chakalova L, Debrand E, Mitchel JA et al (2005) Replication and transcription: shaping the landscape of the genome. *Nature Rev Genet* 6:669–678.

Cremer T, Cremer M, Dietzel S et al (2006) Chromosome territories—a functional nuclear landscape. *Curr Opin Cell Biol* 18:307–316.

Ebert A, Lein S, Schotta G & Reuter G (2006) Histone modification and the control of heterochromatic gene silencing in *Drosophila*. *Chromosome Res* 14:377–392.

Fraser P & Bickmore W (2007) Nuclear organization of the genome and the potential for gene regulation. *Nature* 447:413–417.

Handwerger KE & Gall JG (2006) Subnuclear organelles: new insights into form and function. *Trends Cell Biol* 16:19–26.

Hirano T (2006) At the heart of the chromosome: SMC proteins in action. *Nature Rev Mol Cell Biol* 7:311–322.

Lamond AI & Spector DL (2003) Nuclear speckles: a model for nuclear organelles. *Nature Rev Mol Cell Biol* 4:605–612.

Maeshima K & Laemmli UK (2003) A two-step scaffolding model for mitotic chromosome assembly. *Dev Cell* 4:467–480.

Sims JK, Houston SI, Magazinnik T & Rice JC (2006) A trans-tail histone code defined by monomethylated H4 Lys-20 and H3 Lys-9 demarcates distinct regions of silent chromatin. *J Biol Chem* 281:12760–12766.

Speicher MR & Carter NP (2005) The new cytogenetics: blurring the boundaries with molecular biology. *Nature Rev Genet* 6:782–792.

Zhimulev IF (1998) Morphology and structure of polytene chromosomes. *Adv Genet* 37:1–566.

How Genomes Evolve

Batzer MA & Deininger PL (2002) ALU repeats and human genomic diversity. *Nature Rev Genet* 3:370–379.

Blanchette M, Green ED, Miller W, & Haussler D (2004) Reconstructing large regions of an ancestral mammalian genome in silico. *Genome Res* 14:2412–2423.

Cheng Z, Ventura M, She X et al (2005) A genome-wide comparison of recent chimpanzee and human segmental duplications. *Nature* 437:88–93.

Feuk L, Carson AR & Scherer S (2006) Structural variation in the human genome. *Nature Rev Genet* 7:85–97.

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

International Human Genome Consortium (2004) Finishing the euchromatic sequence of the human genome. *Nature* 431:931–945.

Koszul R, Caburet S, Dujon B & Fischer G (2004) Eucaryotic genome evolution through the spontaneous duplication of large chromosomal segments. *EMBO J* 23:234–243.

Margulies EH, NISC Comparative Sequencing Program & Green ED (2003) Detecting highly conserved regions of the human genome by multispecies sequence comparisons. *Cold Spring Harbor Symp Quant Biol* 68:255–263.

Mouse Genome Sequencing Consortium (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* 420:520–562.

Pollard KS, Salama SR, Lambert N et al (2006) An RNA gene expressed during cortical development evolved rapidly in humans. *Nature* 443:167–172.

Sharp AJ, Cheng Z & Eichler EE (2007) Structural variation of the human genome. *Annu Rev Genomics Hum Genet* 7:407–442.

Siepel A, Bejerano G, Pedersen JS et al (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* 15:1034–1050.

The International HapMap Consortium (2005) A haplotype map of the human genome. *Nature* 437:1299–1320.

The ENCODE Project Consortium (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447:799–816.

DNA Replication, Repair, and Recombination

5

The ability of cells to maintain a high degree of order in a chaotic universe depends upon the accurate duplication of vast quantities of genetic information carried in chemical form as DNA. This process, called *DNA replication*, must occur before a cell can produce two genetically identical daughter cells. Maintaining order also requires the continued surveillance and repair of this genetic information because DNA inside cells is repeatedly damaged by chemicals and radiation from the environment, as well as by thermal accidents and reactive molecules generated inside the cell. In this chapter, we describe the protein machines that replicate and repair the cell's DNA. These machines catalyze some of the most rapid and accurate processes that take place within cells, and their mechanisms clearly demonstrate the elegance and efficiency of cell chemistry.

While the short-term survival of a cell can depend on preventing changes in its DNA, the long-term survival of a species requires that DNA sequences be changeable over many generations. Despite the great efforts that cells make to protect their DNA, occasional changes in DNA sequences do occur. Over time, these changes provide the genetic variation upon which selection pressures act during the evolution of organisms.

We begin this chapter with a brief discussion of the changes that occur in DNA as it is passed down from generation to generation. Next, we discuss the cell mechanisms—DNA replication and DNA repair—that are responsible for minimizing these changes. Finally, we consider some of the most intriguing pathways that alter DNA sequences—those of DNA recombination, including the movement of special DNA sequences in chromosomes called transposable elements.

In This Chapter

THE MAINTENANCE OF DNA SEQUENCES 263

DNA REPLICATION MECHANISMS 266

THE INITIATION AND COMPLETION OF DNA REPLICATION IN CHROMOSOMES 281

DNA REPAIR 295

HOMOLOGOUS RECOMBINATION 304

TRANSPOSITION AND CONSERVATIVE SITE-SPECIFIC RECOMBINATION 316

THE MAINTENANCE OF DNA SEQUENCES

Although, as just pointed out, occasional genetic changes enhance the long-term survival of a species, the survival of the individual demands a high degree of genetic stability. Only rarely do the cell's DNA-maintenance processes fail, resulting in permanent change in the DNA. Such a change is called a **mutation**, and it can destroy an organism if it occurs in a vital position in the DNA sequence.

Mutation Rates Are Extremely Low

The **mutation rate**, the rate at which observable changes occur in DNA sequences, can be determined directly from experiments carried out with a bacterium such as *Escherichia coli*—a resident of our intestinal tract and a commonly used laboratory organism (discussed in Chapter 1). Under laboratory conditions, *E. coli* divides about once every 40 minutes, and a single cell can generate a very large population—several billion—in less than a day. In such a population, it is possible to detect the small fraction of bacteria that have suffered a

damaging mutation in a particular gene, if that gene is not required for the bacterium's survival. For example, the mutation rate of a gene specifically required for cells to use the sugar lactose as an energy source can be determined when the cells are grown in the presence of a different sugar, such as glucose. The fraction of damaged genes underestimates the actual mutation rate because many mutations are *silent* (for example, those that change a codon but not the amino acid it specifies, or those that change an amino acid without affecting the activity of the protein coded for by the gene). After correcting for these silent mutations, one finds that a single gene that encodes an average-sized protein (~10^3 coding nucleotide pairs) accumulates a mutation (not necessarily one that would inactivate the protein) about once in about 10^6 bacterial cell generations. Stated differently, bacteria display a mutation rate of about 1 nucleotide change per 10^9 nucleotides per cell generation.

Recently it has become possible to measure the germ line mutation rate directly in more complex, sexually reproducing organisms such as the nemotode *C. elegans*. These worms, whose generation time is 4 days, were grown for many generations using their self-fertilization mode of reproduction (discussed in Chapter 22). The DNA sequence of a large region of the genome was then determined for many different descendent worms and compared with that of the progenitor worm. This analysis showed, on average, two new mutations (mostly short insertions and deletions) arise in the haploid genome each generation. When the number of cell divisions needed to produce sperm and eggs is taken into account, the mutation rate is roughly 1 mutation per 10^9 nucleotides per cell division, a rate remarkably similar to that in the asexually reproducing *E. coli* described above.

Direct measurement of the germ-line mutation rate in mammals is more difficult, but indirect estimates can be obtained. One way is to compare the amino acid sequences of the same protein in several species. The fraction of the amino acids that differ between any two species can then be compared with the estimated number of years since that pair of species diverged from a common ancestor, as determined from the fossil record. Using this method, one can calculate the number of years that elapse, on average, before an inherited change in the amino acid sequence of a protein becomes fixed in an organism. Because each such change usually reflects the alteration of a single nucleotide in the DNA sequence of the gene encoding that protein, we can use this value to estimate the average number of years required to produce a single, stable mutation in the gene.

These calculations will nearly always substantially underestimate the actual mutation rate, because many mutations will spoil the function of the protein and vanish from the population because of natural selection—that is, by the preferential death of the organisms that contain them. But the sequence of one family of protein fragments does not seem to matter, allowing the genes that encode them to accumulate mutations without being selected against. These are the *fibrinopeptides*, fragments twenty amino acids long that are discarded when the protein *fibrinogen* is activated to form *fibrin* during blood clotting. Since the function of fibrinopeptides apparently does not depend on their amino acid sequence, fibrinopeptides can tolerate almost any amino acid change. Sequence comparisons of the fibrinopeptides can therefore be used to estimate the mutation rate in the germ line. As determined from these studies, a typical protein of 400 amino acids will suffer an amino acid alteration roughly once every 200,000 years.

Another way to estimate mutation rates in humans is to use DNA sequencing to compare corresponding nucleotide sequences directly from closely related species in regions of the genome that do not appear to carry critical information. As expected, such comparisons produce estimates of the mutation rate that agree with those obtained from the fibrinopeptide studies.

E. coli, worms and humans differ greatly in their modes of reproduction and in their generation times. Yet, when the mutation rates of each are normalized to a single round of DNA replication, they are found to be similar: roughly 1 nucleotide change per 10^9 nucleotides each time that DNA is replicated.

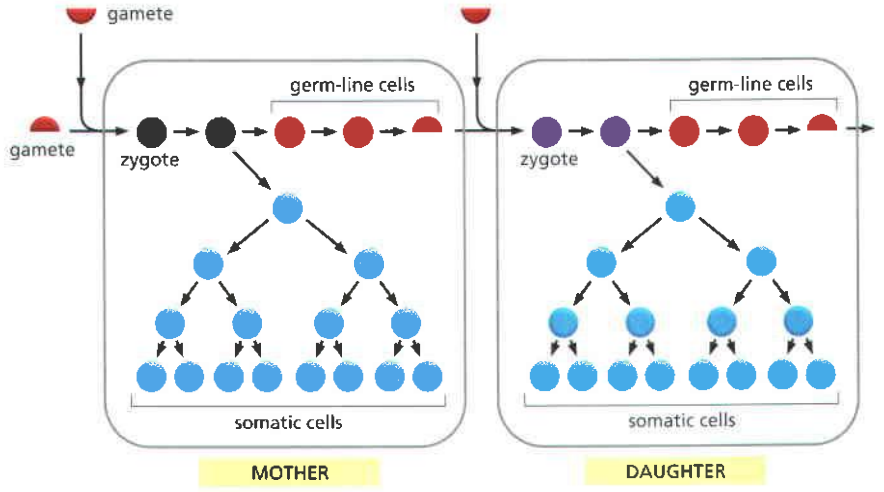


Figure 5–1 Germ-line cells and somatic cells carry out fundamentally different functions. In sexually reproducing organisms, the germ-line cells *(red)* propagate genetic information into the next generation. Somatic cells *(blue)*, which form the body of the organism, are necessary for the survival of germ-line cells but do not themselves leave any progeny.

Low Mutation Rates Are Necessary for Life as We Know It

Since many mutations are deleterious, no species can afford to allow them to accumulate at a high rate in its germ cells. Although the observed mutation frequency is low, it is nevertheless thought to limit the number of essential proteins that any organism can encode to perhaps 50,000. By an extension of the same argument, a mutation frequency tenfold higher would limit an organism to about 5000 essential genes. In this case, evolution would have been limited to organisms considerably less complex than a fruit fly.

The cells of a sexually reproducing organism are of two types: **germ cells** and **somatic cells**. The germ cells transmit genetic information from parent to offspring; the somatic cells form the body of the organism (Figure 5–1). We have seen that germ cells must be protected against high rates of mutation to maintain the species. However, the somatic cells of multicellular organisms must also be protected from genetic change to safeguard each individual. Nucleotide changes in somatic cells can give rise to variant cells, some of which, through natural selection, proliferate rapidly at the expense of the rest of the organism. In an extreme case, the result is an uncontrolled cell proliferation known as cancer, a disease that causes more than 20% of the deaths each year in Europe and North America. These deaths are due largely to an accumulation of changes in the DNA sequences of somatic cells (discussed in Chapter 23). A significant increase in the mutation frequency would presumably cause a disastrous increase in the incidence of cancer by accelerating the rate at which somatic cell variants arise. Thus, both for the perpetuation of a species with a large number of genes (germ cell stability) and for the prevention of cancer resulting from mutations in somatic cells (somatic cell stability), multicellular organisms like ourselves depend on the remarkably high fidelity with which their DNA sequences are replicated and maintained.

Summary

In all cells, DNA sequences are maintained and replicated with high fidelity. The mutation rate, approximately 1 nucleotide change per 10^9 nucleotides each time the DNA is replicated, is roughly the same for organisms as different as bacteria and humans. Because of this remarkable accuracy, the sequence of the human genome (approximately 3×10^9 nucleotide pairs) is changed by only about 3 nucleotides each time a cell divides. This allows most humans to pass accurate genetic instructions from one generation to the next, and also to avoid the changes in somatic cells that lead to cancer.

DNA REPLICATION MECHANISMS

All organisms must duplicate their DNA with extraordinary accuracy before each cell division. In this section, we explore how an elaborate "replication machine" achieves this accuracy, while duplicating DNA at rates as high as 1000 nucleotides per second.

Base-Pairing Underlies DNA Replication and DNA Repair

As introduced in Chapter 1, *DNA templating* is the mechanism the cell uses to copy the nucleotide sequence of one DNA strand into a complementary DNA sequence (Figure 5–2). This process entails the recognition of each nucleotide in the DNA *template strand* by a free (unpolymerized) complementary nucleotide, and it requires the separation of the two strands of the DNA helix. This separation exposes the hydrogen-bond donor and acceptor groups on each DNA base for base-pairing with the appropriate incoming free nucleotide, aligning it for its enzyme-catalyzed polymerization into a new DNA chain.

The first nucleotide-polymerizing enzyme, **DNA polymerase**, was discovered in 1957. The free nucleotides that serve as substrates for this enzyme were found to be deoxyribonucleoside triphosphates, and their polymerization into DNA required a single-stranded DNA template. Figure 5–3 and Figure 5–4 illustrate the stepwise mechanism of this reaction.

The DNA Replication Fork Is Asymmetrical

During DNA replication inside a cell, each of the two original DNA strands serves as a template for the formation of an entire new strand. Because each of the two daughters of a dividing cell inherits a new DNA double helix containing one original and one new strand (Figure 5–5), the DNA double helix is said to be replicated "semiconservatively" by DNA polymerase. How is this feat accomplished?

Analyses carried out in the early 1960s on whole replicating chromosomes revealed a localized region of replication that moves progressively along the parental DNA double helix. Because of its Y-shaped structure, this active region is called a **replication fork** (Figure 5–6). At the replication fork, a multienzyme complex that contains the DNA polymerase synthesizes the DNA of both new daughter strands.

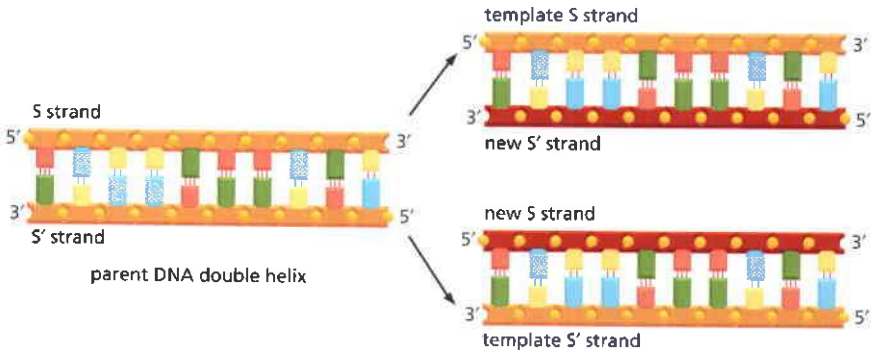


Figure 5–2 **The DNA double helix acts as a template for its own duplication.** Because the nucleotide A will pair successfully only with T, and G only with C, each strand of DNA can serve as a template to specify the sequence of nucleotides in its complementary strand by DNA base-pairing. In this way, a double-helical DNA molecule can be copied precisely.

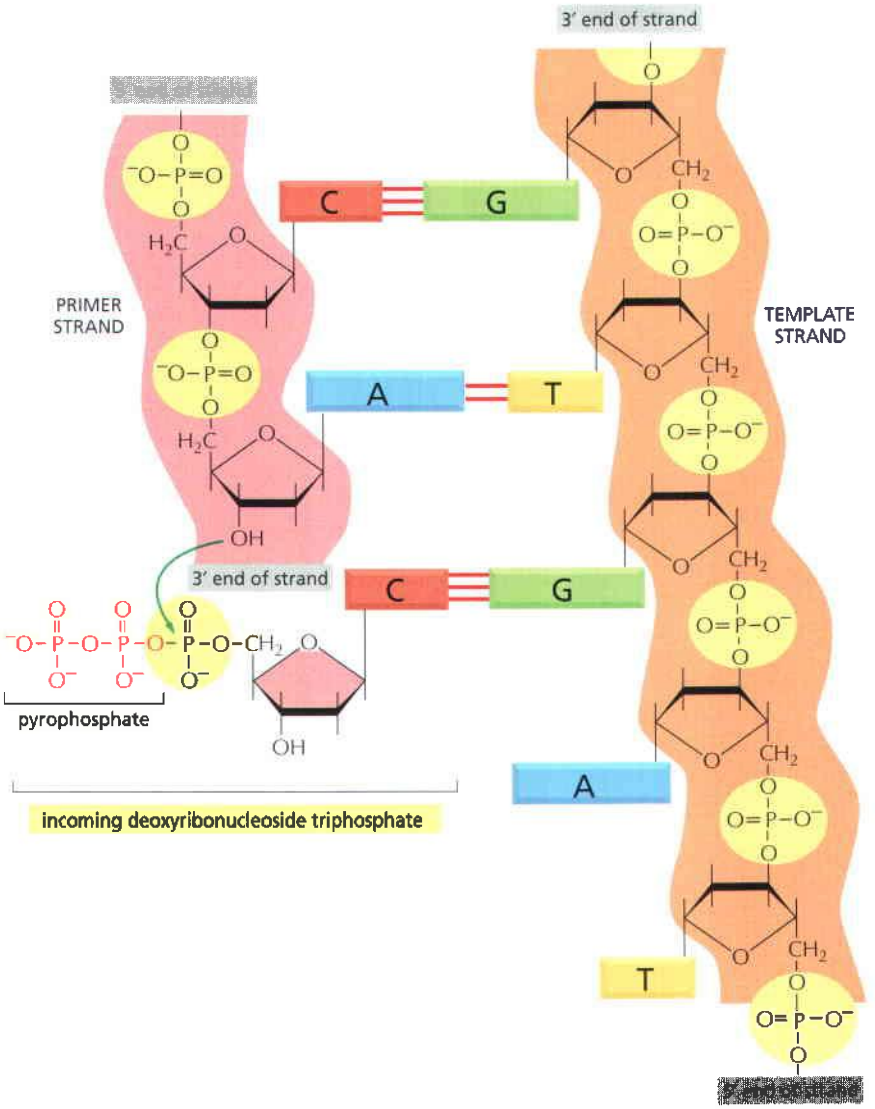


Figure 5–3 The chemistry of DNA synthesis. The addition of a deoxyribonucleotide to the 3′ end of a polynucleotide chain (the *primer strand*) is the fundamental reaction by which DNA is synthesized. As shown, base-pairing between an incoming deoxyribonucleoside triphosphate and an existing strand of DNA (the *template strand*) guides the formation of the new strand of DNA and causes it to have a complementary nucleotide sequence.

Initially, the simplest mechanism of DNA replication seemed to be the continuous growth of both new strands, nucleotide by nucleotide, at the replication fork as it moves from one end of a DNA molecule to the other. But because of the antiparallel orientation of the two DNA strands in the DNA double helix (see Figure 5–2), this mechanism would require one daughter strand to polymerize in the 5′-to-3′ direction and the other in the 3′-to-5′ direction. Such a replication fork would require two distinct types of DNA polymerase enzymes. However, all of the many DNA polymerases that have been discovered can synthesize only in the 5′-to-3′ direction.

How, then, can a DNA strand grow in the 3′-to-5′ direction? The answer was first suggested by the results of an experiment performed in the late 1960s. Researchers added highly radioactive ^{3}H-thymidine to dividing bacteria for a few seconds, so that only the most recently replicated DNA—that just behind the replication fork—became radiolabeled. This experiment revealed the transient existence of pieces of DNA that were 1000–2000 nucleotides long, now

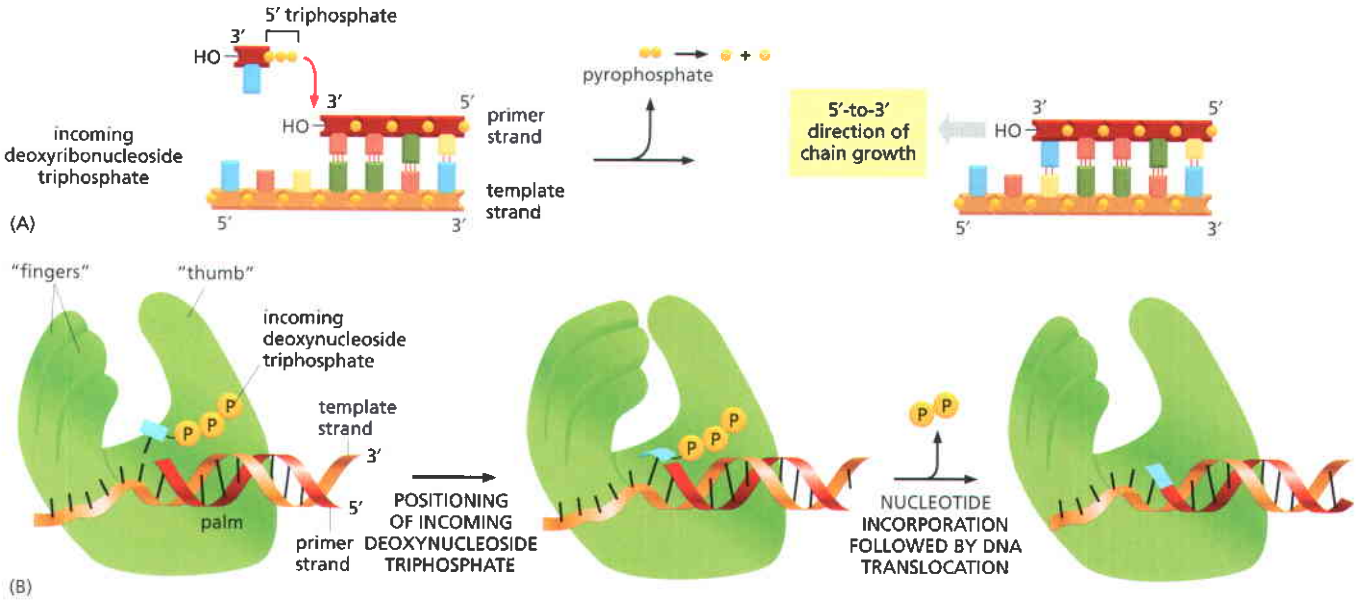


Figure 5–4 DNA synthesis catalyzed by DNA polymerase. (A) As indicated, DNA polymerase catalyzes the stepwise addition of a deoxyribonucleotide to the 3′-OH end of a polynucleotide chain, the *primer strand* that is paired to a second *template strand*. The newly synthesized DNA strand therefore polymerizes in the 5′-to-3′ direction as shown in the previous figure. Because each incoming deoxyribonucleoside triphosphate must pair with the template strand to be recognized by the DNA polymerase, this strand determines which of the four possible deoxyribonucleotides (A, C, G, or T) will be added. The reaction is driven by a large, favorable free-energy change, caused by the release of pyrophosphate and its subsequent hydrolysis to two molecules of inorganic phosphate. (B) The shape of a DNA polymerase molecule, as determined by x-ray crystallography. Roughly speaking DNA polymerases resemble a right hand in which the palm, fingers, and thumb grasp the DNA and form the active site. In the sequence shown, the correct positioning of an incoming deoxynucleoside triphosphate causes the fingers of the polymerase to tighten, thereby initiating the nucleotide addition reaction. Dissociation of pyrophosphate causes release of the fingers and translocation of the DNA by one nucleotide so the active site of the polymerase is ready to receive the next deoxynucleoside triphosphate.

commonly known as *Okazaki fragments*, at the growing replication fork. (Similar replication intermediates were later found in eucaryotes, where they are only 100–200 nucleotides long.) The Okazaki fragments were shown to be polymerized only in the 5′-to-3′ chain direction and to be joined together after their synthesis to create long DNA chains.

A replication fork therefore has an asymmetric structure (**Figure 5–7**). The DNA daughter strand that is synthesized continuously is known as the **leading strand**. Its synthesis slightly precedes the synthesis of the daughter strand that is synthesized discontinuously, known as the **lagging strand**. For the lagging strand, the direction of nucleotide polymerization is opposite to the overall direction of DNA chain growth. The synthesis of this strand by a discontinuous "backstitching" mechanism means that DNA replication requires only the 5′-to-3′ type of DNA polymerase.

The High Fidelity of DNA Replication Requires Several Proofreading Mechanisms

As discussed above, the fidelity of copying DNA during replication is such that only about 1 mistake occurs for every 10^9 nucleotides copied. This fidelity is much higher than one would expect from the accuracy of complementary base-

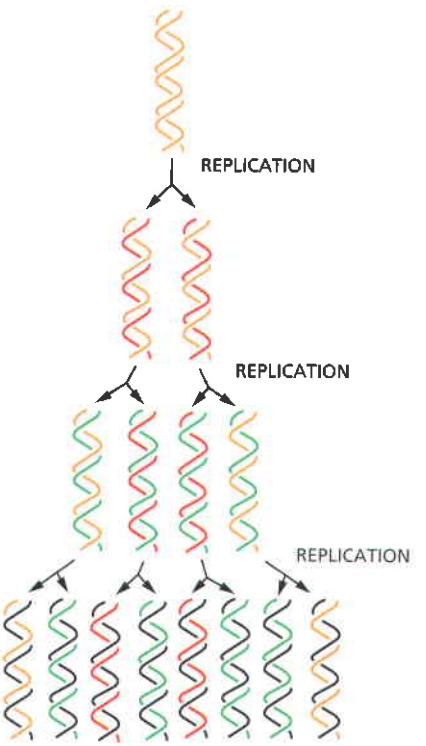


Figure 5–5 The semiconservative nature of DNA replication. In a round of replication, each of the two strands of DNA is used as a template for the formation of a complementary DNA strand. The original strands therefore remain intact through many cell generations.

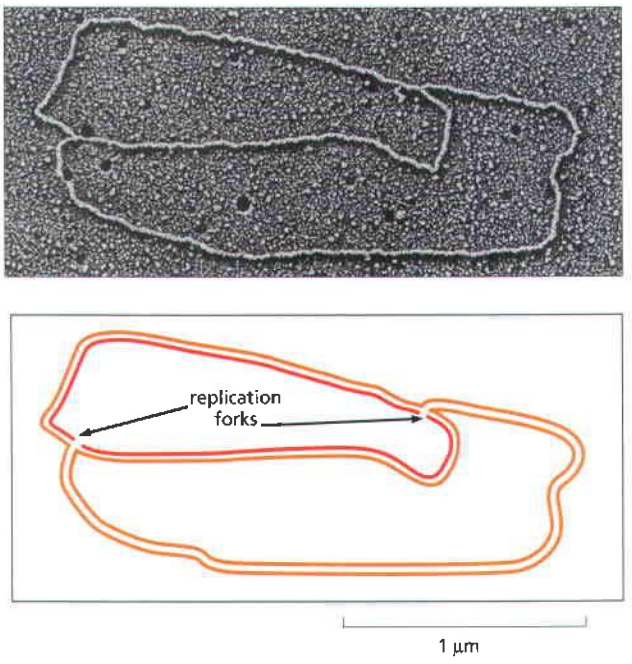


Figure 5–6 Two replication forks moving in opposite directions on a circular chromosome. An active zone of DNA replication moves progressively along a replicating DNA molecule, creating a Y-shaped DNA structure known as a replication fork: the two arms of each Y are the two daughter DNA molecules, and the stem of the Y is the parental DNA helix. In this diagram, parental strands are *orange;* newly synthesized strands are *red.* (Micrograph courtesy of Jerome Vinograd.)

pairing. The standard complementary base pairs (see Figure 4–4) are not the only ones possible. For example, with small changes in helix geometry, two hydrogen bonds can form between G and T in DNA. In addition, rare tautomeric forms of the four DNA bases occur transiently in ratios of 1 part to 10^4 or 10^5. These forms mispair without a change in helix geometry: the rare tautomeric form of C pairs with A instead of G, for example.

If the DNA polymerase did nothing special when a mispairing occurred between an incoming deoxyribonucleoside triphosphate and the DNA template, the wrong nucleotide would often be incorporated into the new DNA chain, producing frequent mutations. The high fidelity of DNA replication, however, depends not only on the initial base-pairing but also on several "proofreading" mechanisms that act sequentially to correct any initial mispairings that might have occurred.

DNA polymerase performs the first proofreading step just before a new nucleotide is added to the growing chain. Our knowledge of this mechanism comes from studies of several different DNA polymerases, including one produced by a bacterial virus, T7, that replicates inside *E. coli.* The correct nucleotide has a higher affinity for the moving polymerase than does the incorrect nucleotide, because the correct pairing is more energetically favorable. Moreover, after nucleotide binding, but before the nucleotide is covalently added to the growing chain, the enzyme must undergo a conformational change in which its "fingers" tighten around the active site (see Figure 5–4). Because this change occurs more readily with correct than incorrect base-pairing, it allows the polymerase to "double-check" the exact base-pair geometry before it catalyzes the addition of the nucleotide.

The next error-correcting reaction, known as *exonucleolytic proofreading,* takes place immediately after those rare instances in which an incorrect nucleotide is covalently added to the growing chain. DNA polymerase enzymes

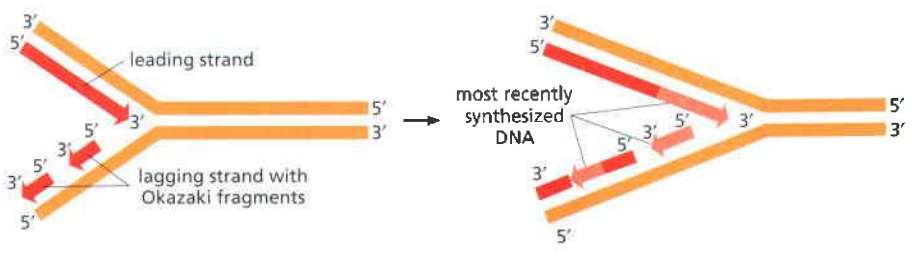


Figure 5–7 The structure of a DNA replication fork. Because both daughter DNA strands are polymerized in the 5′-to-3′ direction, the DNA synthesized on the lagging strand must be made initially as a series of short DNA molecules, called *Okazaki fragments.* On the lagging strand, the Okazaki fragments are synthesized sequentially, with those nearest the fork being the most recently made.

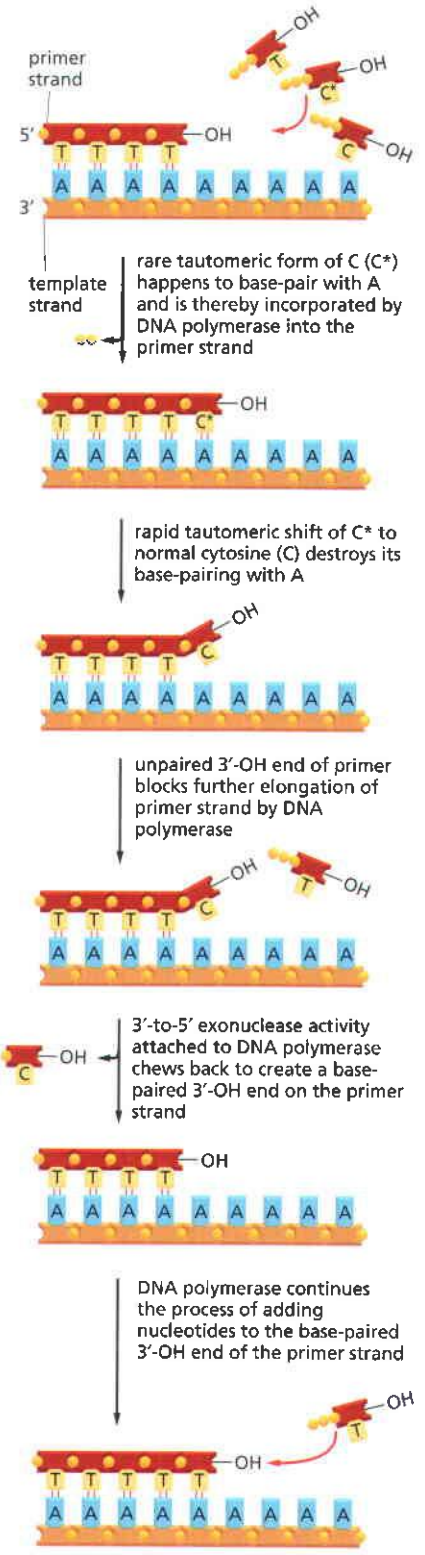


Figure 5–8 Exonucleolytic proofreading by DNA polymerase during DNA replication. In this example, the mismatch is due to the incorporation of a rare, transient tautomeric form of C, indicated by an asterisk. But the same proofreading mechanism applies to any misincorporation at the growing 3′-OH end. The part of DNA polymerase that removes the misincorporated nucleotide is a specialized member of a large class of enzymes, known as *exonucleases,* that cleave nucleotides one at a time from the ends of polynucleotides.

are highly discriminating in the types of DNA chains they will elongate: they absolutely require a previously formed base-paired 3′-OH end of a *primer strand* (see Figure 5–4). Those DNA molecules with a mismatched (improperly base-paired) nucleotide at the 3′-OH end of the primer strand are not effective as templates because the polymerase cannot extend such a strand. DNA polymerase molecules correct such a mismatched primer strand by means of a separate catalytic site (either in a separate subunit or in a separate domain of the polymerase molecule, depending on the polymerase). This *3′-to-5′ proofreading exonuclease* clips off any unpaired residues at the primer terminus, continuing until enough nucleotides have been removed to regenerate a correctly base-paired 3′-OH terminus that can prime DNA synthesis. In this way, DNA polymerase functions as a "self-correcting" enzyme that removes its own polymerization errors as it moves along the DNA (**Figure 5–8** and **Figure 5–9**).

The self-correcting properties of the DNA polymerase depend on its requirement for a perfectly base-paired primer terminus, and it is apparently not possible for such an enzyme to start synthesis *de novo.* By contrast, the RNA polymerase enzymes involved in gene transcription do not need such an efficient exonucleolytic proofreading mechanism: errors in making RNA are not passed on to the next generation, and the occasional defective RNA molecule that is produced has no long-term significance. RNA polymerases are thus able to start new polynucleotide chains without a primer.

There is an error frequency of about 1 mistake for every 10^4 polymerization events both in RNA synthesis and in the separate process of translating mRNA sequences into protein sequences. This error rate is 100,000 times greater than that in DNA replication, where a series of proofreading processes makes the process unusually accurate (**Table 5–1**).

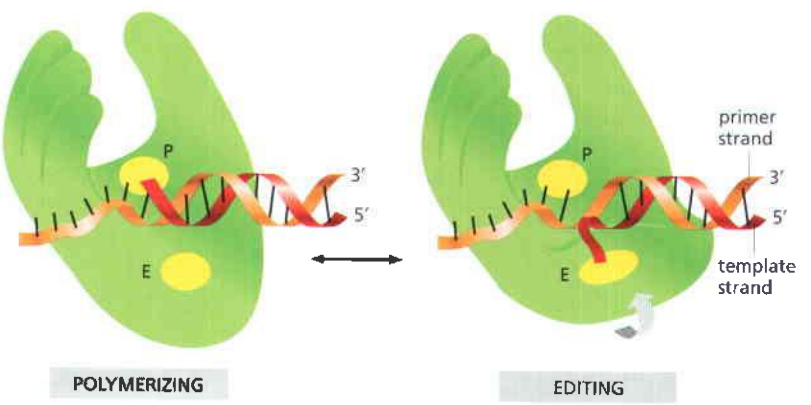


Figure 5–9 Editing by DNA polymerase. Outline of the structures of DNA polymerase complexed with the DNA template in the polymerizing mode *(left)* and the editing mode *(right).* The catalytic site for the exonucleolytic (E) and the polymerization (P) reactions are indicated. In the editing mode, the newly synthesized DNA transiently unpairs from the template and the polymerase undergoes a conformational change, moving the editing catalytic site into place to remove the most recently added nucleotide. <GATT>

Table 5–1 The Three Steps That Give Rise to High-Fidelity DNA Synthesis

REPLICATION STEP	ERRORS PER NUCLEOTIDE
5′ → 3′ polymerization	1 in 10^5
3′ → 5′ exonucleolytic proofreading	1 in 10^2
Strand-directed mismatch repair	1 in 10^2
Combined	1 in 10^9

The third step, strand-directed mismatch repair, is described later in this chapter.

Only DNA Replication in the 5′-to-3′ Direction Allows Efficient Error Correction

The need for accuracy probably explains why DNA replication occurs only in the 5′-to-3′ direction. If there were a DNA polymerase that added deoxyribonucleoside triphosphates in the 3′-to-5′ direction, the growing 5′-chain end, rather than the incoming mononucleotide, would provide the activating triphosphate needed for the covalent linkage. In this case, the mistakes in polymerization could not be simply hydrolyzed away, because the bare 5′-chain end thus created would immediately terminate DNA synthesis (**Figure 5–10**). It is therefore possible to correct a mismatched base only if it has been added to the 3′ end of a DNA chain. Although the backstitching mechanism for DNA replication (see

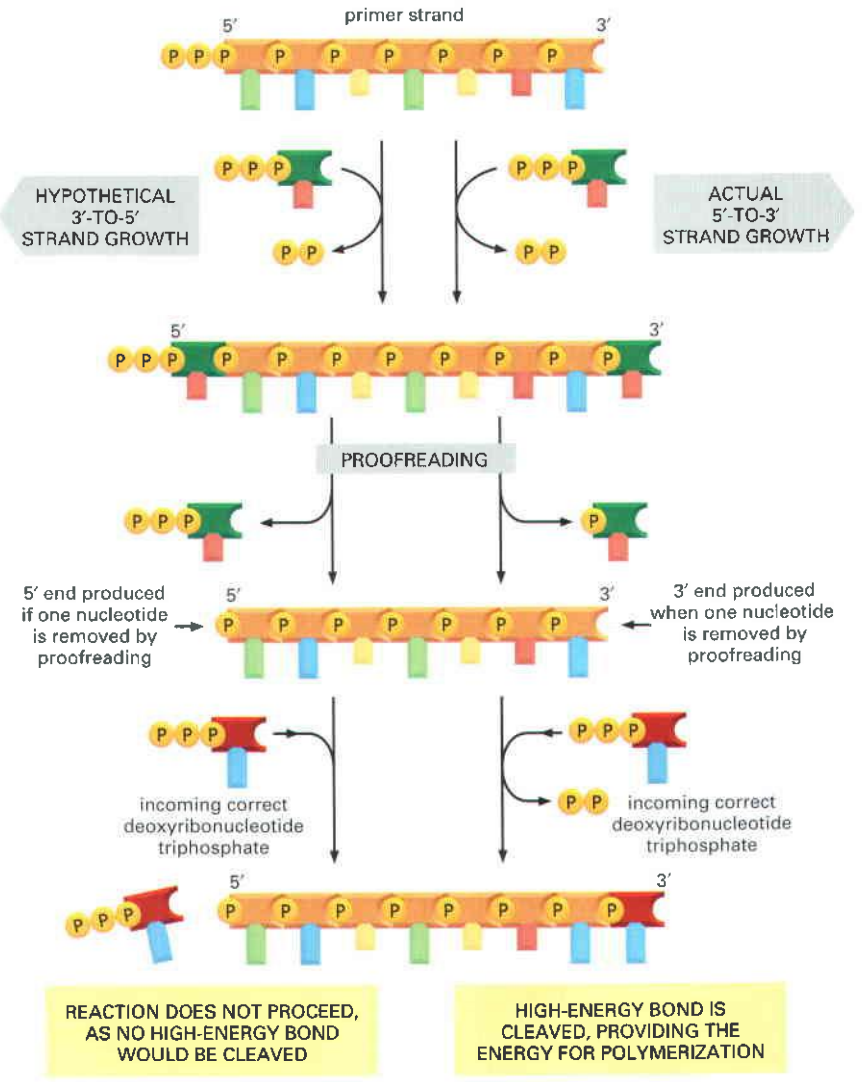


Figure 5–10 An explanation for the 5′-to-3′ direction of DNA chain growth. Growth in the 5′-to-3′ direction, shown on the right, allows the chain to continue to be elongated when a mistake in polymerization has been removed by exonucleolytic proofreading (see Figure 5–8). In contrast, exonucleolytic proofreading in the hypothetical 3′-to-5′ polymerization scheme, shown on the left, would block further chain elongation. For convenience, only the primer strand of the DNA double helix is shown.

Figure 5–7) seems complex, it preserves the 5′-to-3′ direction of polymerization that is required for exonucleolytic proofreading.

Despite these safeguards against DNA replication errors, DNA polymerases occasionally make mistakes. However, as we shall see later, cells have yet another chance to correct these errors by a process called *strand-directed mismatch repair*. Before discussing this mechanism, however, we describe the other types of proteins that function at the replication fork.

A Special Nucleotide-Polymerizing Enzyme Synthesizes Short RNA Primer Molecules on the Lagging Strand

For the leading strand, a special primer is needed only at the start of replication: once a replication fork is established, the DNA polymerase is continuously presented with a base-paired chain end on which to add new nucleotides. On the lagging side of the fork, however, every time the DNA polymerase completes a short DNA Okazaki fragment (which takes a few seconds), it must start synthesizing a completely new fragment at a site further along the template strand (see Figure 5–7). A special mechanism produces the base-paired primer strand required by this DNA polymerase molecule. The mechanism involves an enzyme called **DNA primase**, which uses ribonucleoside triphosphates to synthesize short **RNA primers** on the lagging strand (**Figure 5–11**). In eucaryotes, these primers are about 10 nucleotides long and are made at intervals of 100–200 nucleotides on the lagging strand.

The chemical structure of RNA was introduced in Chapter 1 and is described in detail in Chapter 6. Here, we note only that RNA is very similar in structure to DNA. A strand of RNA can form base pairs with a strand of DNA, generating a DNA/RNA hybrid double helix if the two nucleotide sequences are complementary. Thus the same templating principle used for DNA synthesis guides the synthesis of RNA primers. Because an RNA primer contains a properly base-paired nucleotide with a 3′-OH group at one end, it can be elongated by the DNA polymerase at this end to begin an Okazaki fragment. The synthesis of each Okazaki fragment ends when this DNA polymerase runs into the RNA primer attached to the 5′ end of the previous fragment. To produce a continuous DNA chain from the many DNA fragments made on the lagging strand, a special DNA repair system acts quickly to erase the old RNA primer and replace it with DNA. An enzyme called **DNA ligase** then joins the 3′ end of the new DNA fragment to the 5′ end of the previous one to complete the process (**Figures 5–12 and 5–13**).

Why might an erasable RNA primer be preferred to a DNA primer that would not need to be erased? The argument that a self-correcting polymerase cannot start chains *de novo* also implies its converse: an enzyme that starts chains anew cannot be efficient at self-correction. Thus, any enzyme that primes the synthesis of Okazaki fragments will of necessity make a relatively inaccurate copy (at least 1 error in 10^5). Even if the copies retained in the final product constituted as little as 5% of the total genome (for example, 10 nucleotides per 200-nucleotide DNA fragment), the resulting increase in the overall mutation rate would be enormous. It therefore seems likely that the use of RNA rather than DNA for priming brings a powerful advantage to the cell: the ribonucleotides in the primer automatically mark these sequences as "suspect copy" to be efficiently removed and replaced.

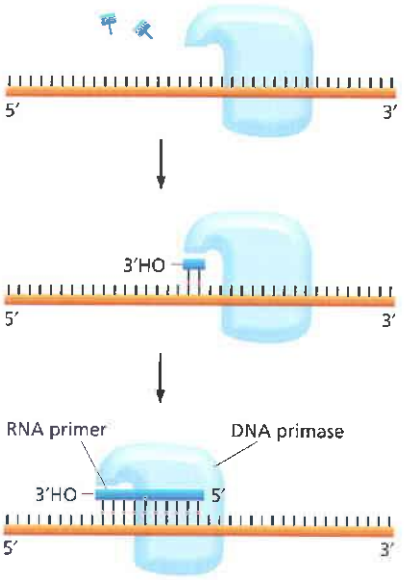


Figure 5–11 RNA primer synthesis. A schematic view of the reaction catalyzed by *DNA primase*, the enzyme that synthesizes the short RNA primers made on the lagging strand using DNA as a template. Unlike DNA polymerase, this enzyme can start a new polynucleotide chain by joining two nucleoside triphosphates together. The primase synthesizes a short polynucleotide in the 5′-to-3′ direction and then stops, making the 3′ end of this primer available for the DNA polymerase.

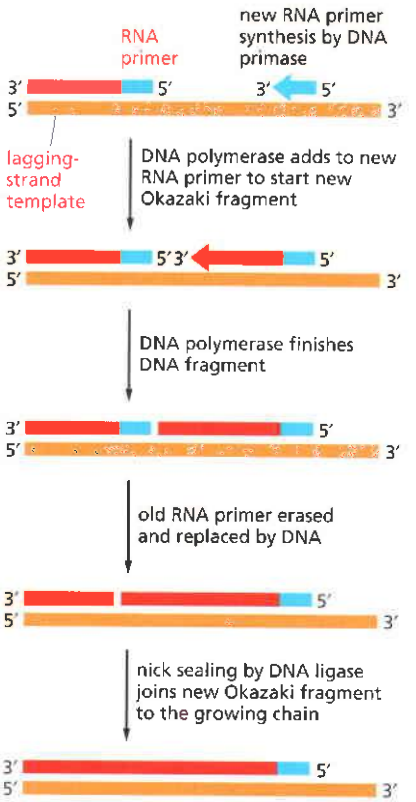


Figure 5–12 The synthesis of one of many DNA fragments on the lagging strand. In eucaryotes, RNA primers are made at intervals spaced by about 200 nucleotides on the lagging strand, and each RNA primer is approximately 10 nucleotides long. This primer is erased by a special DNA repair enzyme (an RNAse H) that recognizes an RNA strand in an RNA/DNA helix and fragments it; this leaves gaps that are filled in by DNA polymerase and DNA ligase.

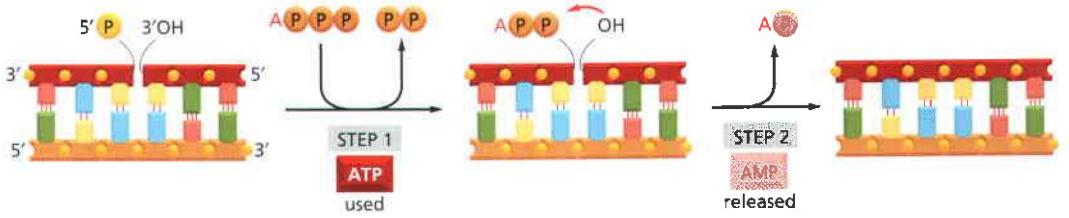


Figure 5–13 **The reaction catalyzed by DNA ligase.** This enzyme seals a broken phosphodiester bond. As shown, DNA ligase uses a molecule of ATP to activate the 5′ end at the nick (step 1) before forming the new bond (step 2). In this way, the energetically unfavorable nick-sealing reaction is driven by being coupled to the energetically favorable process of ATP hydrolysis.

Special Proteins Help to Open Up the DNA Double Helix in Front of the Replication Fork

For DNA synthesis to proceed, the DNA double helix must be opened up ahead of the replication fork so that the incoming deoxyribonucleoside triphosphates can form base pairs with the template strand. However, the DNA double helix is very stable under physiological conditions; the base pairs are locked in place so strongly that it requires temperatures approaching that of boiling water to separate the two strands in a test tube. For this reason, two additional types of replication proteins—DNA helicases and single-strand DNA-binding proteins—are needed to open the double helix and thus provide the appropriate single-stranded DNA template for the DNA polymerase to copy.

DNA helicases were first isolated as proteins that hydrolyze ATP when they are bound to single strands of DNA. As described in Chapter 3, the hydrolysis of ATP can change the shape of a protein molecule in a cyclical manner that allows the protein to perform mechanical work. DNA helicases use this principle to propel themselves rapidly along a DNA single strand. When they encounter a region of double helix, they continue to move along their strand, thereby prying apart the helix at rates of up to 1000 nucleotide pairs per second (**Figures 5–14 and 5–15**).

The two strands of DNA have opposite polarities, and, in principle, a helicase could unwind the DNA double helix by moving in the 5′ to 3′ direction along one strand or in the 3′ to 5′ direction along the other. In fact, both types of DNA helicase exist. In the best understood replication systems in bacteria, a helicase moving 5′ to 3′ along the lagging-strand template appears to have the predominant role, for reasons that will become clear shortly.

Single-strand DNA-binding (SSB) proteins, also called *helix-destabilizing proteins*, bind tightly and cooperatively to exposed single-stranded DNA without covering the bases, which therefore remain available for templating. These proteins are unable to open a long DNA helix directly, but they aid helicases by stabilizing the unwound, single-stranded conformation. In addition, their cooperative binding coats and straightens out the regions of single-stranded DNA on the lagging-strand template, thereby preventing the formation of the short hairpin helices that readily form in single-strand DNA (**Figures 5–16 and 5–17**). These hairpin helices can impede the DNA synthesis catalyzed by DNA polymerase.

A Sliding Ring Holds a Moving DNA Polymerase onto the DNA

On their own, most DNA polymerase molecules will synthesize only a short string of nucleotides before falling off the DNA template. The tendency to dissociate

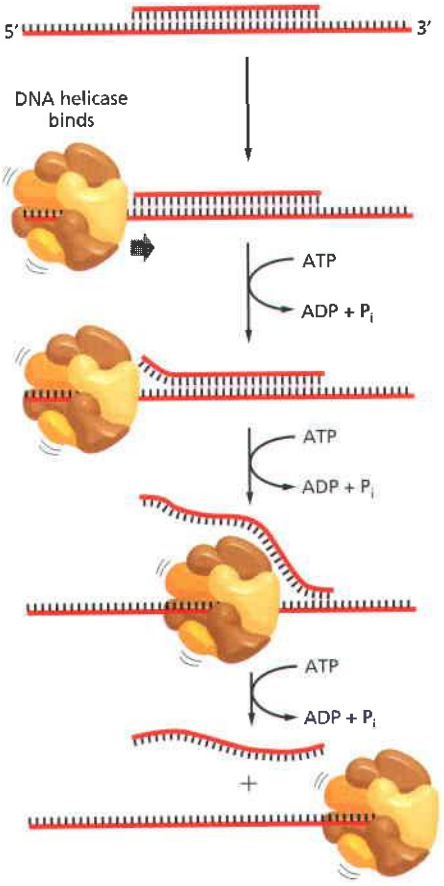


Figure 5–14 **An assay used to test for DNA helicase enzymes.** A short DNA fragment is annealed to a long DNA single strand to form a region of DNA double helix. The double helix is melted as the helicase runs along the DNA single strand, releasing the short DNA fragment in a reaction that requires the presence of both the helicase protein and ATP. The rapid stepwise movement of the helicase is powered by its ATP hydrolysis (see Figure 3–77). As indicated, many DNA helicases are composed of six subunits.

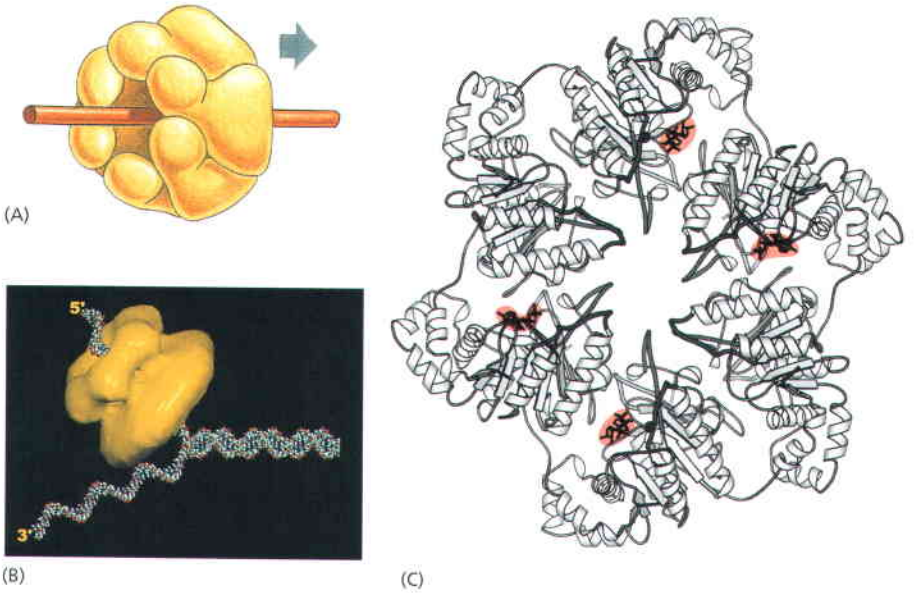


Figure 5–15 The structure of a DNA helicase. <TGCC> (A) A schematic diagram of the protein as a hexameric ring. (B) Schematic diagram showing a DNA replication fork and helicase to scale. (C) Detailed structure of the bacteriophage T7 replicative helicase, as determined by x-ray diffraction. Six identical subunits bind and hydrolyze ATP in an ordered fashion to propel this molecule like a rotary engine along a DNA single strand that passes through the central hole. *Red* indicates bound ATP molecules in the structure. (B, courtesy of Edward H. Egelman; C, from M.R. Singleton et al., *Cell* 101:589–600, 2000. With permission from Elsevier.)

quickly from a DNA molecule allows a DNA polymerase molecule that has just finished synthesizing one Okazaki fragment on the lagging strand to be recycled quickly, so as to begin the synthesis of the next Okazaki fragment on the same strand. This rapid dissociation, however, would make it difficult for the polymerase to synthesize the long DNA strands produced at a replication fork were it not for an accessory protein that functions as a regulated **sliding clamp**. This clamp keeps the polymerase firmly on the DNA when it is moving, but releases it as soon as the polymerase runs into a double-stranded region of DNA.

How can a sliding clamp prevent the polymerase from dissociating without at the same time impeding the polymerase's rapid movement along the DNA molecule? The three-dimensional structure of the clamp protein, determined by x-ray diffraction, reveals that it forms a large ring around the DNA double helix. One side of the ring binds to the back of the DNA polymerase, and the whole ring slides freely along the DNA as the polymerase moves. The assembly of the clamp around the DNA requires ATP hydrolysis by a special protein complex, the **clamp loader**, which hydrolyzes ATP as it loads the clamp on to a primer–template junction (**Figure 5–18**).

On the leading-strand template, the moving DNA polymerase is tightly bound to the clamp, and the two remain associated for a very long time. The DNA polymerase on the lagging-strand template also makes use of the clamp, but each time the polymerase reaches the 5′ end of the preceding Okazaki fragment, the polymerase releases itself from the clamp and dissociates from the

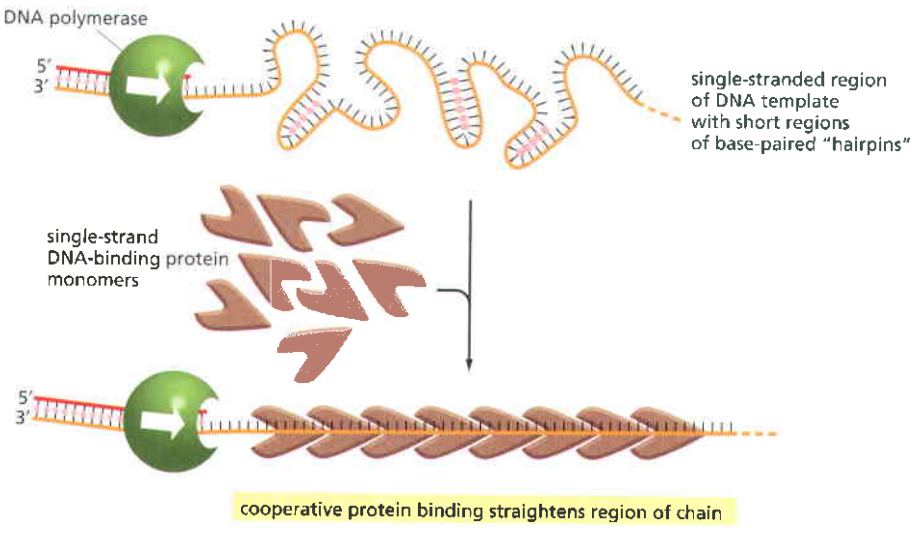


Figure 5–16 The effect of single-strand DNA-binding proteins (SSB proteins) on the structure of single-stranded DNA. Because each protein molecule prefers to bind next to a previously bound molecule, long rows of this protein form on a DNA single strand. This *cooperative binding* straightens out the DNA template and facilitates the DNA polymerization process. The "hairpin helices" shown in the bare, single-stranded DNA result from a chance matching of short regions of complementary nucleotide sequence; they are similar to the short helices that typically form in RNA molecules (see Figure 1–6).

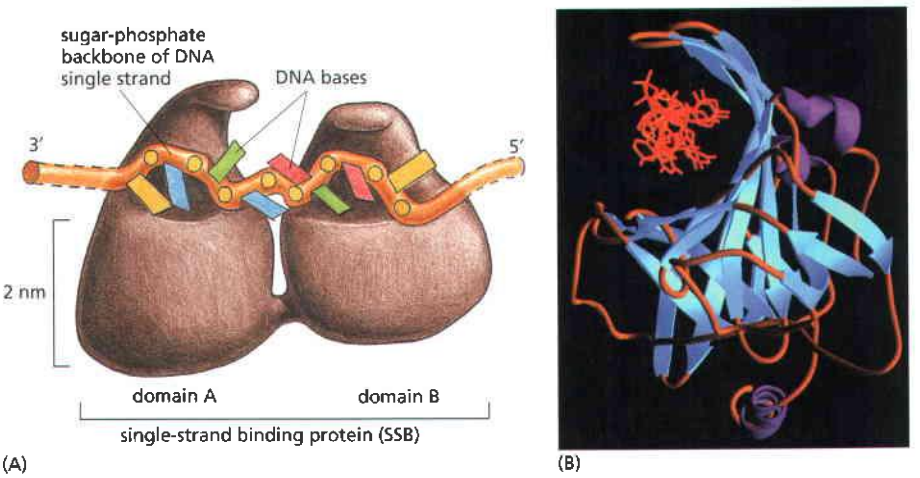
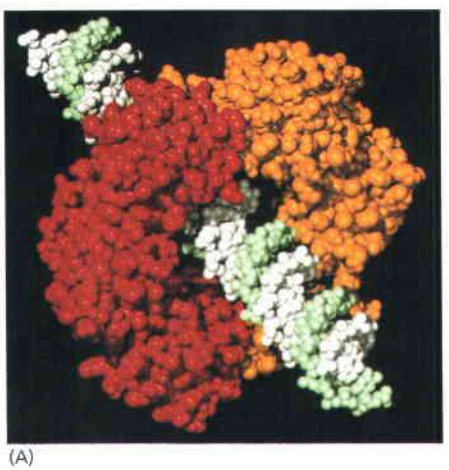


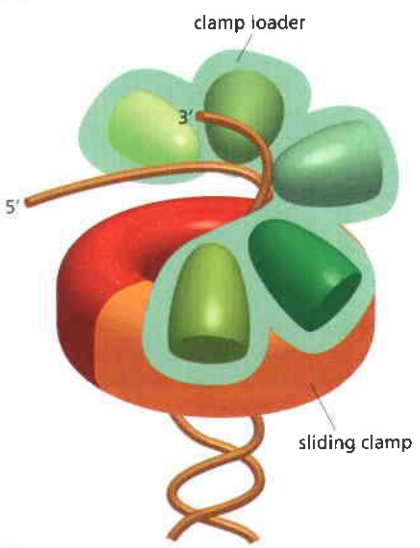
Figure 5–17 The structure of the human single-strand binding protein bound to DNA. (A) A front view of the two DNA-binding domains of RPA protein, which cover a total of eight nucleotides. Note that the DNA bases remain exposed in this protein–DNA complex. (B) A diagram showing the three-dimensional structure, with the DNA strand *(red)* viewed end-on. (B, from A. Bochkarev et al., *Nature* 385:176–181, 1997. With permission from Macmillan Publishers Ltd.)

template. This polymerase molecule then associates with a new clamp that is assembled on the RNA primer of the next Okazaki fragment.

The Proteins at a Replication Fork Cooperate to Form a Replication Machine

Although we have discussed DNA replication as though it were performed by a mix of proteins all acting independently, in reality most of the proteins are held together in a large and orderly multienzyme complex that rapidly synthesizes





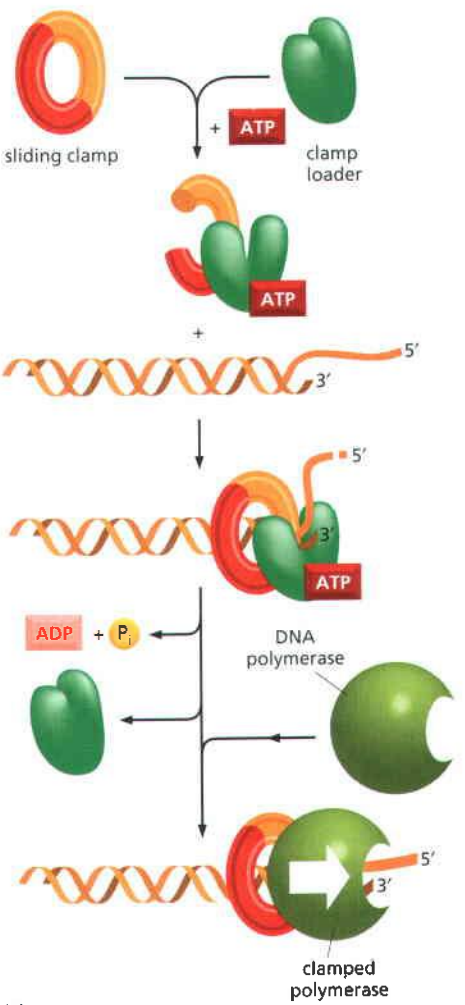
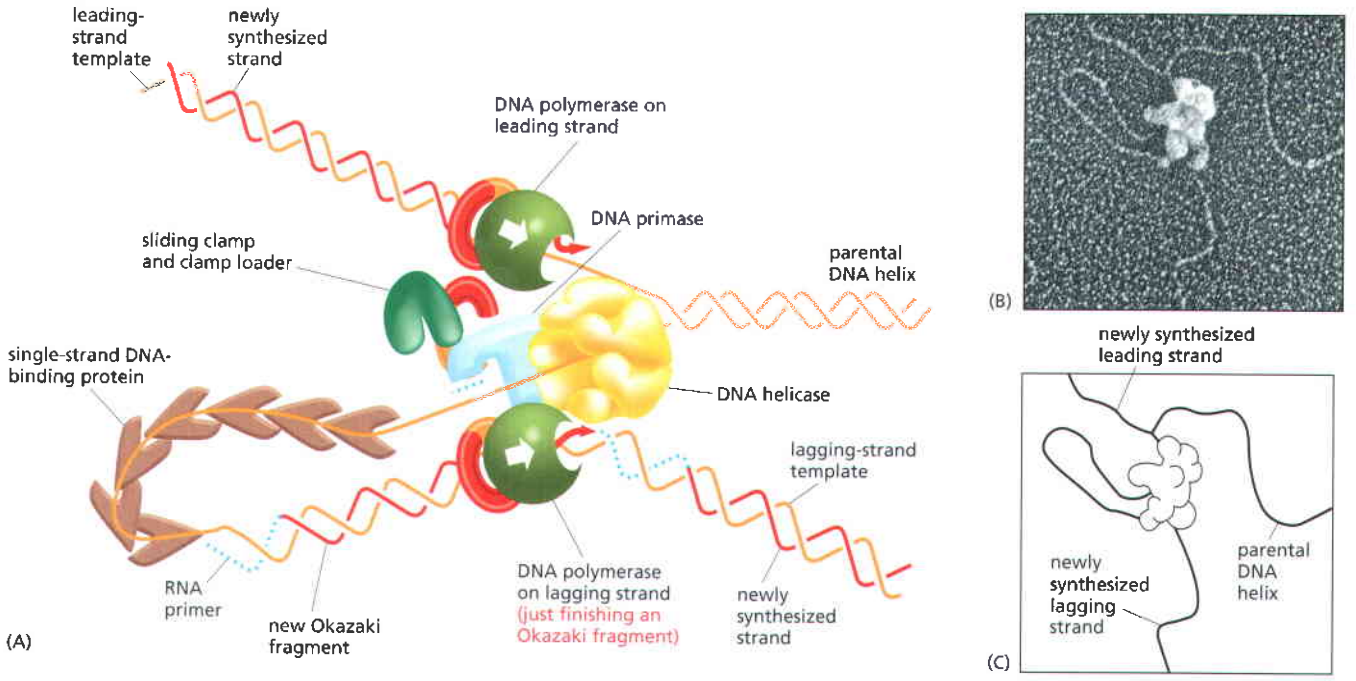


Figure 5–18 The regulated sliding clamp that holds DNA polymerase on the DNA. (A) The structure of the clamp protein from *E. coli*, as determined by X-ray crystallography, with a DNA helix added to indicate how the protein fits around DNA. (B) The structure of the five-subunit clamp loader resembles a screw nut, with its threads matching the grooves of DNA. It appears to tighten around the primer junction until its further progress is blocked by the 3′ end of the primer, at which point the loader hydrolyzes ATP and releases the clamp. (C) Schematic illustration showing how the clamp is assembled to hold a moving DNA polymerase molecule on the DNA. In the simplified reaction shown here, the clamp loader dissociates into solution once the clamp has been assembled. At a true replication fork, the clamp loader remains close to the lagging-strand polymerase, ready to assemble a new clamp at the start of each new Okazaki fragment (see Figure 5–19). (A, from X.P. Kong et al., *Cell* 69:425–437, 1992. With permission from Elsevier; C, from G.D. Bowman, M. O'Donnell and J. Kuriyan, *Nature* 429:708–709, 2004. With permission from Macmillan Publishers Ltd.)



DNA. This complex can be likened to a tiny sewing machine composed of protein parts and powered by nucleoside triphosphate hydrolyses. Like a sewing machine, the replication complex probably remains stationary with respect to its immediate surroundings; the DNA can be thought of as a long strip of cloth being rapidly threaded through it. Although the replication complex has been most intensively studied in *E. coli* and several of its viruses, a very similar complex also operates in eucaryotes, as we see below.

Figure 5–19A summarizes the functions of the subunits of the replication machine. At the front of the replication fork, DNA helicase opens the DNA helix. Two DNA polymerase molecules work at the fork, one on the leading strand and one on the lagging strand. Whereas the DNA polymerase molecule on the leading strand can operate in a continuous fashion, the DNA polymerase molecule on the lagging strand must restart at short intervals, using a short RNA primer made by a DNA primase molecule. The close association of all these protein components increases the efficiency of replication and is made possible by a folding back of the lagging strand as shown in Figure 5–19A. This arrangement also facilitates the loading of the polymerase clamp each time that an Okazaki fragment is synthesized: the clamp loader and the lagging-strand DNA polymerase molecule are kept in place as a part of the protein machine even when they detach from their DNA template. The replication proteins are thus linked together into a single large unit (total molecular weight $>10^6$ daltons), enabling DNA to be synthesized on both sides of the replication fork in a coordinated and efficient manner.

On the lagging strand, the DNA replication machine leaves behind a series of unsealed Okazaki fragments, which still contain the RNA that primed their synthesis at their 5′ ends. This RNA is removed and the resulting gap is filled in by DNA repair enzymes that operate behind the replication fork (see Figure 5–12).

Figure 5–19 An active replication fork. (A) This schematic diagram shows a current view of the arrangement of replication proteins at a replication fork when DNA is being synthesized. The lagging-strand DNA has been folded to bring the lagging-strand DNA polymerase molecule into a complex with the leading-strand DNA polymerase molecule. This folding also brings the 3′ end of each completed Okazaki fragment close to the start site for the next Okazaki fragment. Because the lagging-strand DNA polymerase molecule remains bound to the rest of the replication proteins, it can be reused to synthesize successive Okazaki fragments. In this diagram, it is about to let go of its completed DNA fragment and move to the RNA primer that will be synthesized nearby, as required to start the next DNA fragment. Additional proteins (not shown) help to hold the different protein components of the fork together, enabling them to function as a well-coordinated protein machine. <AATA> <CCCG> (B) An electron micrograph showing the replication machine from the bacteriophage T4 as it moves along a template synthesizing DNA behind it. (C) An interpretation of the micrograph is given in the sketch: note especially the DNA loop on the lagging strand. Apparently, the replication proteins became partly detached from the very front of the replication fork during the preparation of this sample for electron microscopy. (B, courtesy of Jack Griffith; see P.D. Chastain et al., *J. Biol. Chem.* 278:21276–21285, 2003.)

A Strand-Directed Mismatch Repair System Removes Replication Errors That Escape from the Replication Machine

As stated previously, bacteria such as *E. coli* are capable of dividing once every 40 minutes, making it relatively easy to screen large populations to find a rare mutant cell that is altered in a specific process. One interesting class of mutants contains alterations in so-called *mutator genes*, which greatly increase the rate of spontaneous mutation. Not surprisingly, one such mutant makes a defective

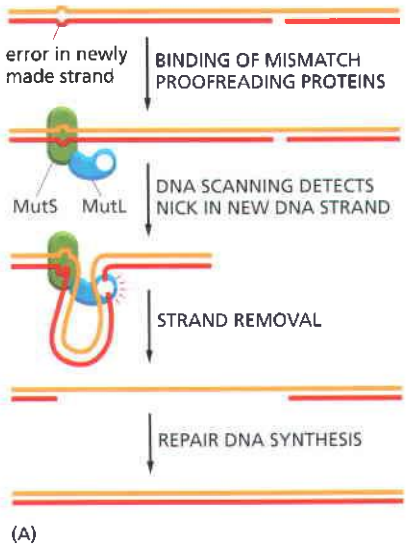
form of the 3′-to-5′ proofreading exonuclease that is a part of the DNA polymerase enzyme (see Figures 5–8 and 5–9). The mutant DNA polymerase no longer proofreads effectively, and many replication errors that would otherwise have been removed accumulate in the DNA.

The study of other *E. coli* mutants exhibiting abnormally high mutation rates has uncovered a proofreading system that removes replication errors made by the polymerase that have been missed by the proofreading exonuclease. This **strand-directed mismatch repair** system detects the potential for distortion in the DNA helix from the misfit between noncomplementary base pairs.

If the proofreading system simply recognized a mismatch in newly replicated DNA and randomly corrected one of the two mismatched nucleotides, it would mistakenly "correct" the original template strand to match the error exactly half the time, thereby failing to lower the overall error rate. To be effective, such a proofreading system must be able to distinguish and remove the mismatched nucleotide only on the newly synthesized strand, where the replication error occurred.

The strand-distinction mechanism used by the mismatch proofreading system in *E. coli* depends on the methylation of selected A residues in the DNA. Methyl groups are added to all A residues in the sequence GATC, but not until some time after the A has been incorporated into a newly synthesized DNA chain. As a result, the only GATC sequences that have not yet been methylated are in the new strands just behind a replication fork. The recognition of these unmethylated GATCs allows the new DNA strands to be transiently distinguished from old ones, as required if their mismatches are to be selectively removed. The three-step process involves recognition of a mismatch, excision of the segment of DNA containing the mismatch from the newly synthesized strand, and resynthesis of the excised segment using the old strand as a template. This strand-directed mismatch repair system reduces the number of errors made during DNA replication by an additional factor of 100 (see Table 5–1, p. 271).

A similar mismatch proofreading system functions in human cells (**Figure 5–20**). The importance of this system in humans is seen in individuals who inherit one defective copy of a mismatch repair gene (along with a functional gene on the other copy of the chromosome). These people have a marked predisposition for certain types of cancers. For example, in a type of colon cancer called *hereditary nonpolyposis colon cancer (HNPCC)*, spontaneous mutation of the remaining functional gene produces a clone of somatic cells that, because they are deficient in mismatch proofreading, accumulate mutations unusually rapidly. Most cancers arise in cells that have accumulated multiple mutations



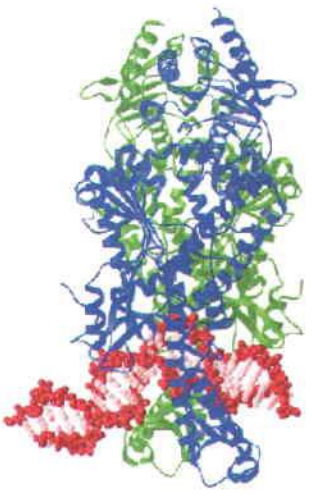


Figure 5–20 A model for strand-directed mismatch repair in eucaryotes. (A) The two proteins shown are present in both bacteria and eucaryotic cells: MutS binds specifically to a mismatched base pair, while MutL scans the nearby DNA for a nick. Once MutL finds a nick, it triggers the degradation of the nicked strand all the way back through the mismatch. Because nicks are largely confined to newly replicated strands in eucaryotes, replication errors are selectively removed. In bacteria, the mechanism is the same, except that an additional protein in the complex (MutH) nicks unmethylated (and therefore newly replicated) GATC sequences, thereby beginning the process illustrated here. (B) The structure of the MutS protein bound to a DNA mismatch. This protein is a dimer, which grips the DNA double helix as shown, kinking the DNA at the mismatched base pair. It seems that the MutS protein scans the DNA for mismatches by testing for sites that can be readily kinked, which are those without a normal complementary base pair. (B, from G. Obmolova et al., *Nature* 407:703–710, 2000. With permission from Macmillan Publishers Ltd.)

(see Figure 20–11), and cells deficient in mismatch proofreading therefore have a greatly enhanced chance of becoming cancerous. Fortunately, most of us inherit two good copies of each gene that encodes a mismatch proofreading protein; this protects us, because it is highly unlikely for both copies to become mutated in the same cell.

In eucaryotes, the mechanism for distinguishing the newly synthesized strand from the parental template strand at the site of a mismatch does not depend on DNA methylation. Indeed, some eucaryotes—including yeasts and *Drosophila*—do not methylate their DNA. Newly synthesized lagging-strand DNA transiently contains *nicks* (before they are sealed by DNA ligase) and biochemical experiments reveal that such nicks (also called *single-strand breaks*) provide the signal that directs the mismatch proofreading system to the appropriate strand (see Figure 5–20). This idea also requires that the newly synthesized DNA on the leading strand be transiently nicked; how this occurs is uncertain.

DNA Topoisomerases Prevent DNA Tangling During Replication

As a replication fork moves along double-stranded DNA, it creates what has been called the "winding problem." Every 10 base pairs replicated at the fork corresponds to one complete turn about the axis of the parental double helix. Therefore, for a replication fork to move, the entire chromosome ahead of the fork would normally have to rotate rapidly (**Figure 5–21**). This would require large amounts of energy for long chromosomes, and an alternative strategy is used instead: a swivel is formed in the DNA helix by proteins known as **DNA topoisomerases**.

A DNA topoisomerase can be viewed as a reversible nuclease that adds itself covalently to a DNA backbone phosphate, thereby breaking a phosphodiester bond in a DNA strand. This reaction is reversible, and the phosphodiester bond re-forms as the protein leaves.

One type of topoisomerase, called *topoisomerase I*, produces a transient single-strand break (or nick); this break in the phosphodiester backbone allows the two sections of DNA helix on either side of the nick to rotate freely relative to each other, using the phosphodiester bond in the strand opposite the nick as a swivel point (**Figure 5–22**). Any tension in the DNA helix will drive this rotation in the direction that relieves the tension. As a result, DNA replication can occur with the rotation of only a short length of helix—the part just ahead of the fork. Because the covalent linkage that joins the DNA topoisomerase protein to a DNA phosphate retains the energy of the cleaved phosphodiester bond, resealing is rapid and does not require additional energy input. In this respect, the rejoining mechanism differs from that catalyzed by the enzyme DNA ligase, discussed previously (see Figure 5–13).

A second type of DNA topoisomerase, *topoisomerase II*, forms a covalent linkage to both strands of the DNA helix at the same time, making a transient *double-strand break* in the helix. These enzymes are activated by sites on chromosomes where two double helices cross over each other. Once a topoisomerase II molecule binds to such a crossing site, the protein uses ATP hydrolysis to perform the following set of reactions efficiently: (1) it breaks one double helix reversibly to create a DNA "gate;" (2) it causes the second, nearby double helix to pass through this break; and (3) it then reseals the break and dissociates from the DNA (**Figure 5–23**). In this way, type II DNA topoisomerases can efficiently separate two interlocked DNA circles (**Figure 5–24**).

The same reaction also prevents the severe DNA tangling problems that would otherwise arise during DNA replication. This role is nicely illustrated by

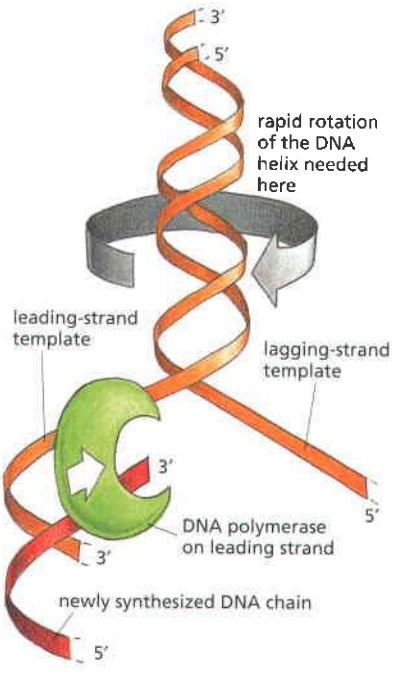


Figure 5–21 The "winding problem" that arises during DNA replication. For a bacterial replication fork moving at 500 nucleotides per second, the parental DNA helix ahead of the fork must rotate at 50 revolutions per second.

mutant yeast cells that produce, in place of the normal topoisomerase II, a version that is inactive above 37°C. When the mutant cells are warmed to this temperature, their daughter chromosomes remain intertwined after DNA replication and are unable to separate. The enormous usefulness of topoisomerase II

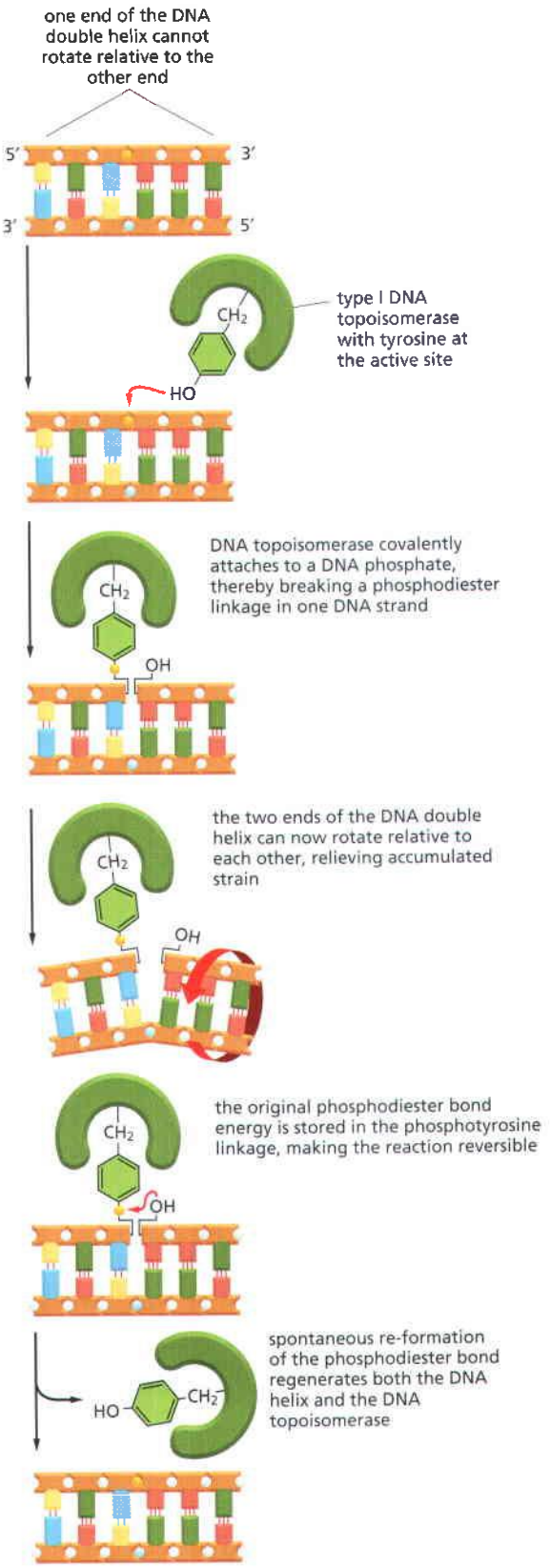


Figure 5–22 The reversible DNA nicking reaction catalyzed by a eucaryotic DNA topoisomerase I enzyme. As indicated, these enzymes transiently form a single covalent bond with DNA; this allows free rotation of the DNA around the covalent backbone bonds linked to the *blue* phosphate.

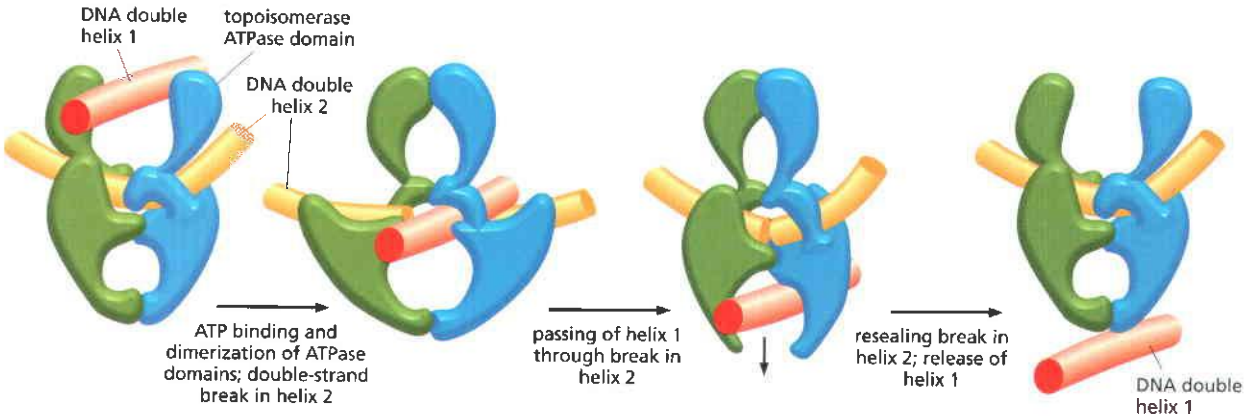


Figure 5–23 A model for topoisomerase II action. As indicated, ATP binding to the two ATPase domains causes them to dimerize and drives the reactions shown. Because a single cycle of this reaction can occur in the presence of a non-hydrolyzable ATP analog, ATP hydrolysis is thought to be needed only to reset the enzyme for each new reaction cycle. This model is based on the structure of enzyme in combination with biochemical experiments. (Modified from J.M. Berger, *Curr. Opin. Struct. Biol.* 8:26–32, 1998. With permission from Elsevier.)

for untangling chromosomes can readily be appreciated by anyone who has struggled to remove a tangle from a fishing line without the aid of scissors.

DNA Replication Is Fundamentally Similar in Eucaryotes and Bacteria

Much of what we know about DNA replication was first derived from studies of purified bacterial and bacteriophage multienzyme systems capable of DNA replication *in vitro*. The development of these systems in the 1970s was greatly facilitated by the prior isolation of mutants in a variety of replication genes; these mutants were exploited to identify and purify the corresponding replication proteins. The first mammalian replication system that accurately replicated DNA *in vitro* was described in the mid-1980s, and mutations in genes encoding nearly all of the replication components have now been isolated and analyzed in the yeast *Saccharomyces cerevisiae*. As a result, much is known about the detailed enzymology of DNA replication in eucaryotes, and it is clear that the fundamental features of DNA replication—including replication fork geometry and the use of a multiprotein replication machine—have been conserved during the long evolutionary process that separated bacteria from eucaryotes.

There are more protein components in eucaryotic replication machines than there are in the bacterial analogs, even though the basic functions are the same. Thus, for example, the eucaryotic single-strand binding (SSB) protein is formed from three subunits, whereas only a single subunit is found in bacteria. Similarly, the eucaryotic DNA primase is incorporated into a multisubunit enzyme that also contains a DNA polymerase called DNA polymerase α-primase. This protein complex begins each Okazaki fragment on the lagging strand with RNA and then extends the RNA primer with a short length of DNA. At this point, the two main eucaryotic replicative polymerases, δ and ε, come into play and complete each Okazaki fragment while simultaneously extending the leading strand. Exactly how the tasks of leading and lagging strand synthesis are distributed between these two DNA polymerases is not yet understood.

As we see in the next section, the eucaryotic replication machinery has the added complication of having to replicate through nucleosomes, the repeating structural unit of chromosomes discussed in Chapter 4. Nucleosomes are spaced at intervals of about 200 nucleotide pairs along the DNA, which may explain why new Okazaki fragments are synthesized on the lagging strand at intervals of 100–200 nucleotides in eucaryotes, instead of 1000–2000 nucleotides as in bacteria. Nucleosomes may also act as barriers that slow down the movement of DNA polymerase molecules, which may be why eucaryotic replication forks move only about one-tenth as fast as bacterial replication forks.

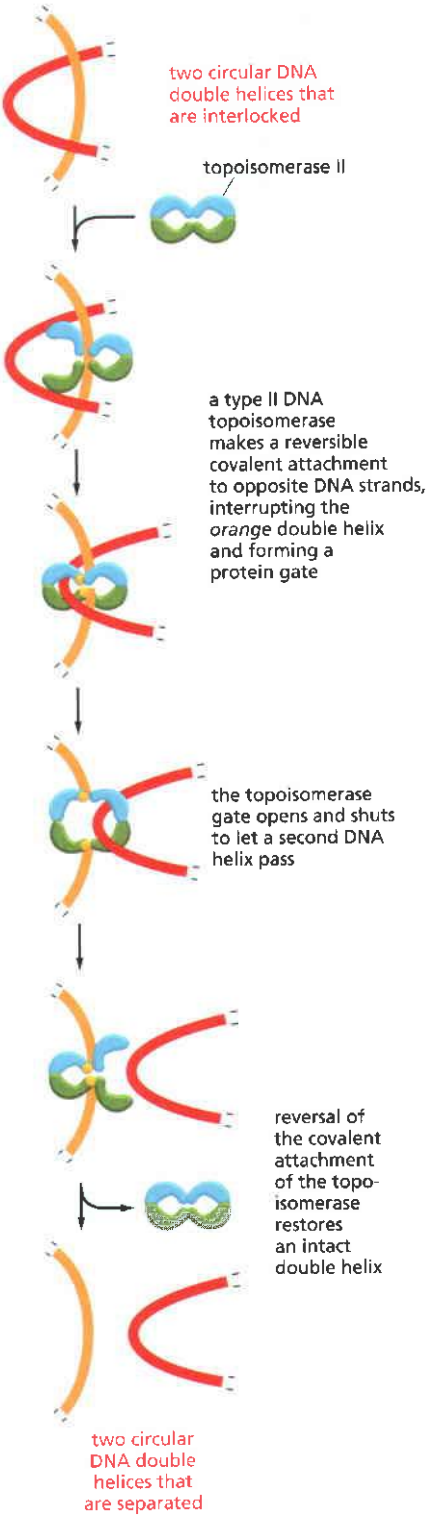


Figure 5–24 The DNA-helix-passing reaction catalyzed by DNA topoisomerase II. Identical reactions are used to untangle DNA inside the cell. Unlike type I topoisomerases, type II enzymes use ATP hydrolysis and some of the bacterial versions can introduce superhelical tension into DNA. Type II topoisomerases are largely confined to proliferating cells in eucaryotes; partly for that reason, they have been popular targets for anticancer drugs.

Summary

DNA replication takes place at a Y-shaped structure called a replication fork. A self-correcting DNA polymerase enzyme catalyzes nucleotide polymerization in a 5′-to-3′ direction, copying a DNA template strand with remarkable fidelity. Since the two strands of a DNA double helix are antiparallel, this 5′-to-3′ DNA synthesis can take place continuously on only one of the strands at a replication fork (the leading strand). On the lagging strand, short DNA fragments must be made by a "backstitching" process. Because the self-correcting DNA polymerase cannot start a new chain, these lagging-strand DNA fragments are primed by short RNA primer molecules that are subsequently erased and replaced with DNA.

DNA replication requires the cooperation of many proteins. These include (1) DNA polymerase and DNA primase to catalyze nucleoside triphosphate polymerization; (2) DNA helicases and single-strand DNA-binding (SSB) proteins to help in opening up the DNA helix so that it can be copied; (3) DNA ligase and an enzyme that degrades RNA primers to seal together the discontinuously synthesized lagging-strand DNA fragments; and (4) DNA topoisomerases to help to relieve helical winding and DNA tangling problems. Many of these proteins associate with each other at a replication fork to form a highly efficient "replication machine," through which the activities and spatial movements of the individual components are coordinated.

THE INITIATION AND COMPLETION OF DNA REPLICATION IN CHROMOSOMES

We have seen how a set of replication proteins rapidly and accurately generates two daughter DNA double helices behind a replication fork. But how is this replication machinery assembled in the first place, and how are replication forks created on a double-stranded DNA molecule? In this section, we discuss how cells initiate DNA replication and how they carefully regulate this process to ensure that it takes place not only at the proper positions on the chromosome but also at the appropriate time in the life of the cell. We also discuss a few of the special problems that the replication machinery in eucaryotic cells must overcome. These include the need to replicate the enormously long DNA molecules found in eucaryotic chromosomes, as well as the difficulty of copying DNA molecules that are tightly complexed with histones in nucleosomes.

DNA Synthesis Begins at Replication Origins

As discussed previously, the DNA double helix is normally very stable: the two DNA strands are locked together firmly by many hydrogen bonds formed between the bases on each strand. To be used as a template, the double helix must be opened up and the two strands separated to expose unpaired bases. As we shall see, the process of DNA replication is begun by special *initiator proteins* that bind to double-stranded DNA and pry the two strands apart, breaking the hydrogen bonds between the bases.

The positions at which the DNA helix is first opened are called **replication origins** (**Figure 5–25**). In simple cells like those of bacteria or yeast, origins are specified by DNA sequences several hundred nucleotide pairs in length. This DNA contains both short sequences that attract initiator proteins and stretches of DNA that are especially easy to open. We saw in Figure 4–4 that an A-T base

pair is held together by fewer hydrogen bonds than a G-C base pair. Therefore, DNA rich in A-T base pairs is relatively easy to pull apart, and regions of DNA enriched in A-T pairs are typically found at replication origins.

Although the basic process of replication fork initiation depicted in Figure 5–25 is fundamentally the same for bacteria and eucaryotes, the detailed way in which this process is performed and regulated differs between these two groups of organisms. We first consider the simpler and better-understood case in bacteria and then turn to the more complex situation found in yeasts, mammals, and other eucaryotes.

Bacterial Chromosomes Typically Have a Single Origin of DNA Replication

The genome of *E. coli* is contained in a single circular DNA molecule of 4.6×10^6 nucleotide pairs. DNA replication begins at a single origin of replication, and the two replication forks assembled there proceed (at approximately 500–1000 nucleotides per second) in opposite directions until they meet up roughly halfway around the chromosome (**Figure 5–26**). The only point at which *E. coli* can control DNA replication is initiation: once the forks have been assembled at the origin, they synthesize DNA at relatively constant speed until replication is finished. Therefore, it is not surprising that the initiation of DNA replication is highly regulated. The process begins when initiator proteins bind in multiple copies to specific sites in the replication origin, wrapping the DNA around the proteins to form a large protein–DNA complex. This complex then attracts a DNA helicase bound to a helix loader, and the helicase is placed around an adjacent DNA single strand whose bases have been exposed by the assembly of the initiator protein–DNA complex. The helicase loader is analogous to the clamp loader we encountered above; it has the additional job of keeping the helicase in an inactive form until it is properly loaded onto a nascent replication fork. Once the helicase is loaded, it begins to unwind DNA, exposing enough single-stranded DNA for primase to synthesize the RNA primer that begins the leading strand (**Figure 5–27**). This quickly leads to the assembly of the remaining proteins to create two replication forks, with protein complexes that move, with respect to the replication origin, in opposite directions. These protein machines continue to synthesize DNA until all of the DNA template downstream of each fork has been replicated.

In *E. coli*, the interaction of the initiator protein with the replication origin is carefully regulated, with initiation occurring only when sufficient nutrients are available for the bacterium to complete an entire round of replication. Not only is the activity of the initiator protein controlled, but an origin of replication that has just been used experiences a "refractory period," caused by a delay in the methylation of newly synthesized A nucleotides. Further initiation of replication is blocked until these As are methylated (**Figure 5–28**).

Eucaryotic Chromosomes Contain Multiple Origins of Replication

We have seen how two replication forks begin at a single replication origin in bacteria and proceed in opposite directions, moving away from the origin until all of the DNA in the single circular chromosome is replicated. The bacterial genome is sufficiently small for these two replication forks to duplicate the genome in about 40 minutes. Because of the much greater size of most eucaryotic chromosomes, a different strategy is required to allow their replication in a timely manner.

A method for determining the general pattern of eucaryotic chromosome replication was developed in the early 1960s. Human cells growing in culture are labeled for a short time with ^{3}H-thymidine so that the DNA synthesized during this period becomes highly radioactive. The cells are then gently lysed, and the DNA is streaked on the surface of a glass slide coated with a photographic emulsion. Development of the emulsion reveals the pattern of labeled DNA through

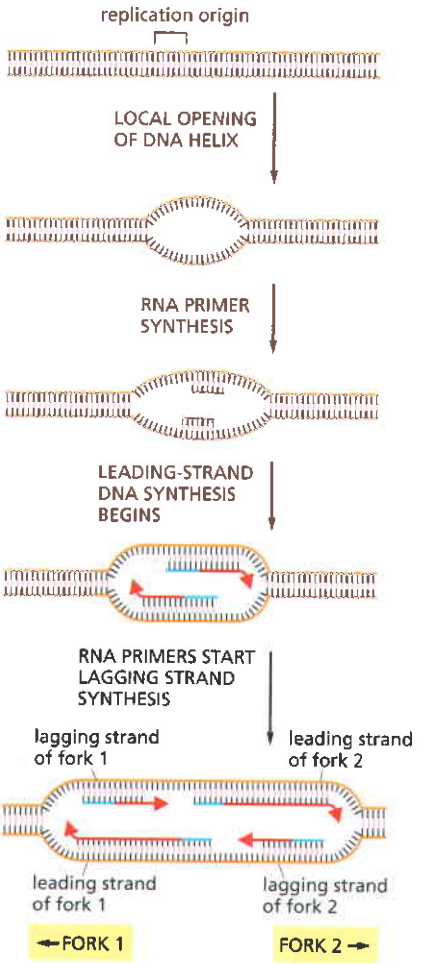


Figure 5–25 A replication bubble formed by replication fork initiation. This diagram outlines the major steps in the initiation of replication forks at replication origins. The structure formed at the last step, in which both strands of the parental DNA helix have been separated from each other and serve as templates for DNA synthesis, is called a *replication bubble*.

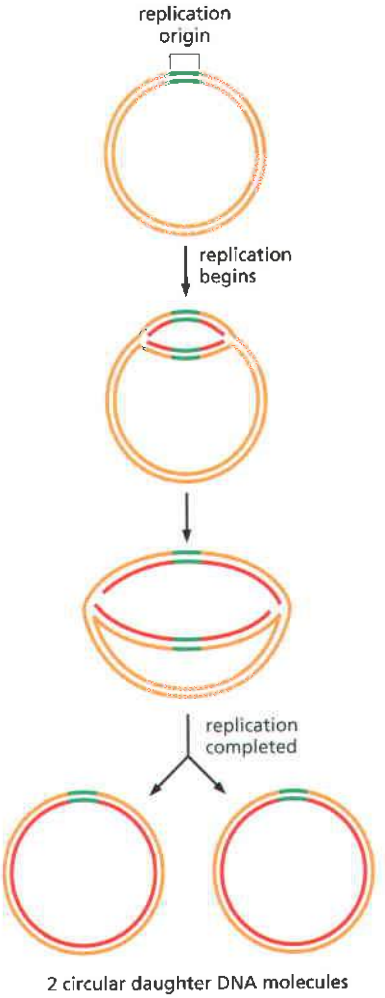


Figure 5–26 DNA replication of a bacterial genome. It takes *E. coli* about 40 minutes to duplicate its genome of 4.6×10^6 nucleotide pairs. For simplicity, no Okazaki fragments are shown on the lagging strand. What happens as the two replication forks approach each other and collide at the end of the replication cycle is not well understood, although the replication machines are disassembled as part of the process.

a technique known as *autoradiography*. The time allotted for radioactive labeling is chosen to allow each replication fork to move several micrometers along the DNA, so that the replicated DNA can be detected in the light microscope as lines of silver grains, even though the DNA molecule itself is too thin to be visible. In this way, both the rate and the direction of replication-fork movement can be determined (**Figure 5–29**). From the rate at which tracks of replicated DNA increase in length with increasing labeling time, the replication forks are estimated to travel at about 50 nucleotides per second. This is approximately one-tenth of the rate at which bacterial replication forks move, possibly reflecting the increased difficulty of replicating DNA that is packaged tightly in chromatin.

An average-size human chromosome contains a single linear DNA molecule of about 150 million nucleotide pairs. It would take 0.02 seconds/nucleotide × 150×10^6 nucleotides = 3.0×10^6 seconds (about 800 hours) to replicate such a DNA molecule from end to end with a single replication fork moving at a rate of 50 nucleotides per second. As expected, therefore, the autoradiographic experiments just described reveal that many forks are moving simultaneously on each eucaryotic chromosome.

Further experiments of this type have shown the following: (1) Replication origins tend to be activated in clusters, called *replication units*, of perhaps 20–80

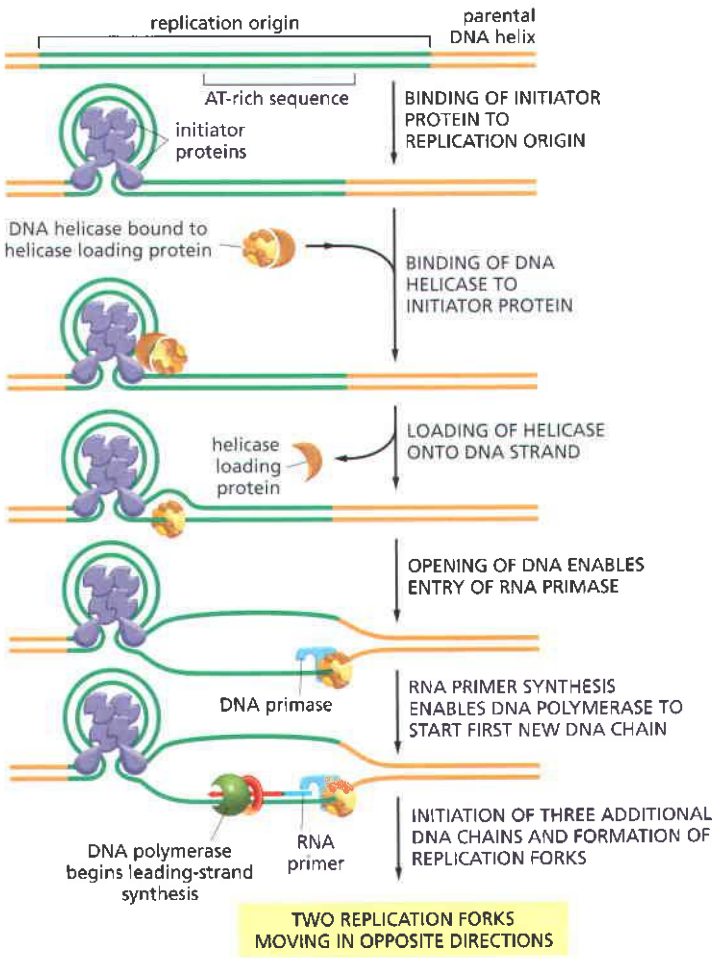


Figure 5–27 The proteins that initiate DNA replication in bacteria. The mechanism shown was established by studies *in vitro* with mixtures of highly purified proteins. For *E. coli* DNA replication, the major initiator protein, the helicase, and the primase are the DnaA, DnaB, and DnaG proteins, respectively. In the first step, several molecules of the initiator protein bind to specific DNA sequences in the replication origin and form a compact structure in which the DNA is wrapped around the protein. Next, the helicase is brought in by a helicase loading protein (the DnaC protein), which inhibits the helicase until it is properly loaded at the replication origin. The helicase loading protein thereby prevents the helicase from inappropriately entering other single-stranded stretches of DNA in the bacterial genome. Aided by single-strand binding protein (not shown), the loaded helicase opens up the DNA thereby enabling primase to enter and synthesize the primer for the first DNA chain. Subsequent steps (not shown) result in the initiation of three additional DNA chains and the final assembly of two complete replication forks.

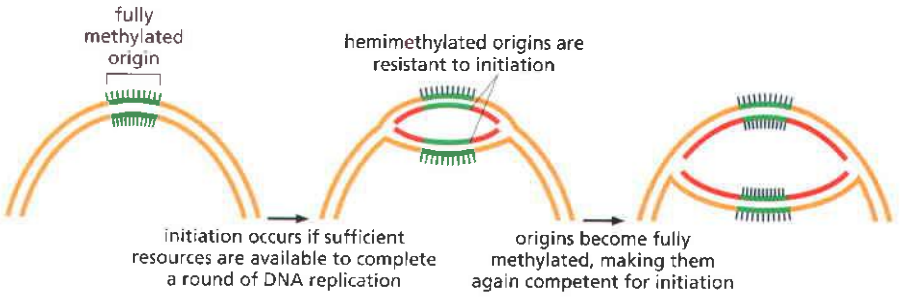


Figure 5–28 Methylation of the *E. coli* replication origin creates a refractory period for DNA initiation. DNA methylation occurs at GATC sequences, 11 of which are found in the origin of replication (spanning approximately 250 nucleotide pairs). In its hemimethylated state, the origin of replication is bound by an inhibitor protein (Seq A, not shown), which blocks access of the origin to initiator proteins. Eventually (about 20 minutes after replication is initiated), the hemimethylated origins become fully methylated by a DNA methylase enzyme; Seq A then dissociates.

A single enzyme, the *Dam* methylase, is responsible for methylating all *E. coli* GATC sequences. A lag in methylation after the replication of GATC sequences is also used by the *E. coli* mismatch proofreading system to distinguish the newly synthesized DNA strand from the parental DNA strand; in that case, the relevant GATC sequences are scattered throughout the chromosome, and they are not bound by Seq A.

origins. (2) New replication units seem to be activated at different times during the cell cycle until all of the DNA is replicated, a point that we return to below. (3) Within a replication unit, individual origins are spaced at intervals of 30,000–250,000 nucleotide pairs from one another. (4) As in bacteria, replication forks are formed in pairs and create a replication bubble as they move in opposite directions away from a common point of origin, stopping only when they collide head-on with a replication fork moving in the opposite direction (or when they reach a chromosome end). In this way, many replication forks operate independently on each chromosome and yet form two complete daughter DNA helices.

In Eucaryotes DNA Replication Takes Place During Only One Part of the Cell Cycle

When growing rapidly, bacteria replicate their DNA continually, and they can begin a new round before the previous one is complete. In contrast, DNA replication in most eucaryotic cells occurs only during a specific part of the cell division cycle, called the *DNA synthesis phase* or **S phase** (Figure 5–30). In a mammalian cell, the S phase typically lasts for about 8 hours; in simpler eucaryotic cells such as yeasts, the S phase can be as short as 40 minutes. By its end, each chromosome has been replicated to produce two complete copies, which remain joined together at their centromeres until the *M phase* (M for *mitosis*), which soon follows. In Chapter 17, we describe the control system that runs the cell cycle and explain why entry into each phase of the cycle requires the cell to have successfully completed the previous phase.

In the following sections, we explore how chromosome replication is coordinated within the S phase of the cell cycle.

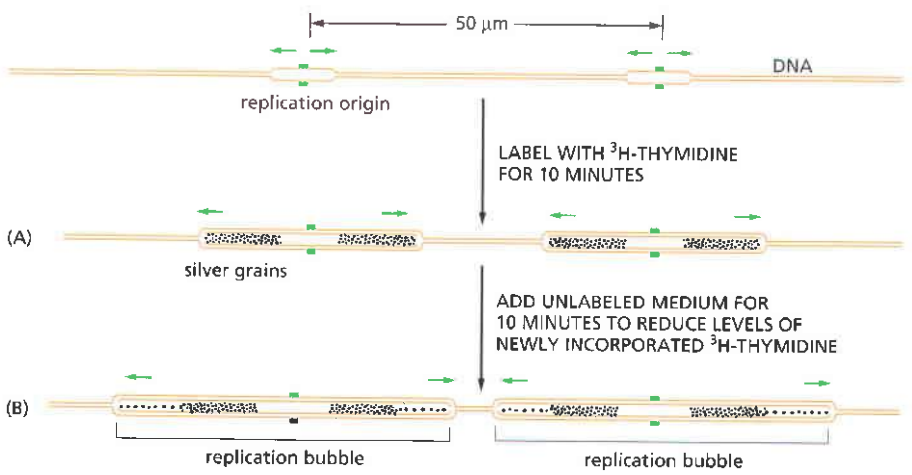


Figure 5–29 The experiments that demonstrated the pattern in which replication forks are formed and move on eucaryotic chromosomes. The new DNA made in human cells in culture was labeled briefly with a pulse of highly radioactive thymidine (^{3}H-thymidine). (A) In this experiment, the cells were lysed, and the DNA was stretched out on a glass slide that was subsequently covered with a photographic emulsion. After several months the emulsion was developed, revealing a line of silver grains over the radioactive DNA. The *brown* DNA in this figure is shown only to help with the interpretation of the autoradiograph; the unlabeled DNA is invisible in such experiments. (B) This experiment was the same except that a further incubation in unlabeled medium allowed additional DNA, with a lower level of radioactivity, to be replicated. The pairs of dark tracks in (B) were found to have silver grains tapering off in opposite directions, demonstrating bidirectional fork movement from a central replication origin where a replication bubble forms (see Figure 5–25). A replication fork is thought to stop only when it encounters a replication fork moving in the opposite direction or when it reaches the end of the chromosome; in this way, all the DNA is eventually replicated.

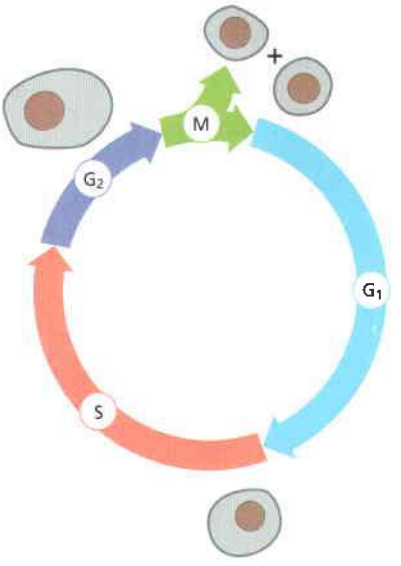


Figure 5–30 The four successive phases of a standard eucaryotic cell cycle. During the G_1, S, and G_2 phases, the cell grows continuously. During M phase growth stops, the nucleus divides, and the cell divides in two. DNA replication is confined to the part of the cell cycle known as S phase. G_1 is the gap between M phase and S phase; G_2 is the gap between S phase and M phase.

Different Regions on the Same Chromosome Replicate at Distinct Times in S Phase

In mammalian cells, the replication of DNA in the region between one replication origin and the next should normally require only about an hour to complete, given the rate at which a replication fork moves and the largest distances measured between replication origins. Yet S phase usually lasts for about 8 hours in a mammalian cell. This implies that the replication origins are not all activated simultaneously and that the DNA in each replication unit (which, as we noted above, contains a cluster of about 20–80 replication origins) is replicated during only a small part of the total S-phase interval.

Are different replication units activated at random, or are different regions of the genome replicated in a specified order? One way to answer this question is to use the thymidine analog bromodeoxyuridine (BrdU) to label the newly synthesized DNA in synchronized cell populations, adding it for different short periods throughout S phase. Later, during M phase, those regions of the mitotic chromosomes that have incorporated BrdU into their DNA can be recognized by their altered staining properties or by means of anti-BrdU antibodies. The results show that different regions of each chromosome are replicated in a reproducible order during S phase (**Figure 5–31**). Moreover, as one would expect from the clusters of replication forks seen in DNA autoradiographs (see Figure 5–29), the timing of replication is coordinated over large regions of the chromosome.

Much more sophisticated methods now exist for monitoring DNA replication initiation and tracking the movement of DNA replication forks in cells. These approaches use DNA microarrays—grids the size of a postage stamp studded with tens of thousands of fragments of known DNA sequence. As we will see in detail in Chapter 8, each different DNA fragment is placed at a unique position on the microarray, and whole genomes can thereby be represented in an orderly manner. If a DNA sample from a group of cells in S phase is broken up and hybridized to a microarray representing that organism's genome, the amount of each DNA sequence can be determined. Because a segment of a genome that has been replicated will contain twice as much DNA as an unreplicated segment, replication fork initiation and fork movement can be accurately monitored in this way (**Figure 5–32**). Although this method provides much greater precision, it leads to many of the same conclusions reached from the earlier studies.

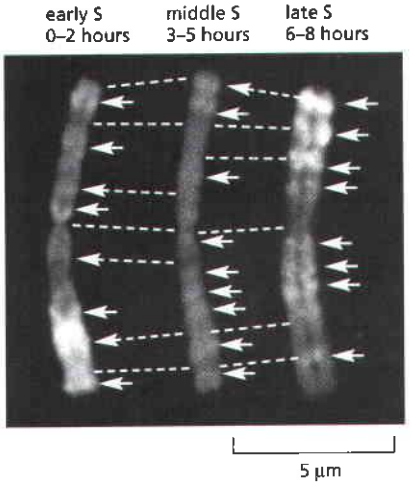


Figure 5–31 Different regions of a chromosome are replicated at different times in S phase. These light micrographs show stained mitotic chromosomes in which the replicating DNA has been differentially labeled during different defined intervals of the preceding S phase. In these experiments, cells were first grown in the presence of BrdU (a thymidine analog) and absence of thymidine to label the DNA uniformly. The cells were then briefly pulsed with thymidine in the absence of BrdU during early, middle, or late S phase. Because the DNA made during the thymidine pulse is a double helix with thymidine on one strand and BrdU on the other, it stains more darkly than the remaining DNA (which has BrdU on both strands) and shows up as a bright band *(arrows)* on these negatives. Broken lines connect corresponding positions on the three identical copies of the chromosome shown. (Courtesy of Elton Stubblefield.)

Highly Condensed Chromatin Replicates Late, While Genes in Less Condensed Chromatin Tend to Replicate Early

It seems that the order in which replication origins are activated depends, in part, on the chromatin structure in which the origins reside. We saw in Chapter 4 that heterochromatin is a particularly condensed state of chromatin, while euchromatin chromatin has a less condensed conformation that is apparently required to allow transcription. Heterochromatin tends to be replicated very late in S phase, suggesting that the timing of replication is related to the packing of the DNA in chromatin. This suggestion is supported by an examination of the two X chromosomes in a female mammalian cell. While these two chromosomes contain essentially the same DNA sequences, one is active for DNA transcription and the other is not (discussed in Chapter 7). Nearly all of the inactive X chromosome is condensed into heterochromatin, and its DNA replicates late in S phase. Its active homolog is less condensed and replicates throughout S phase.

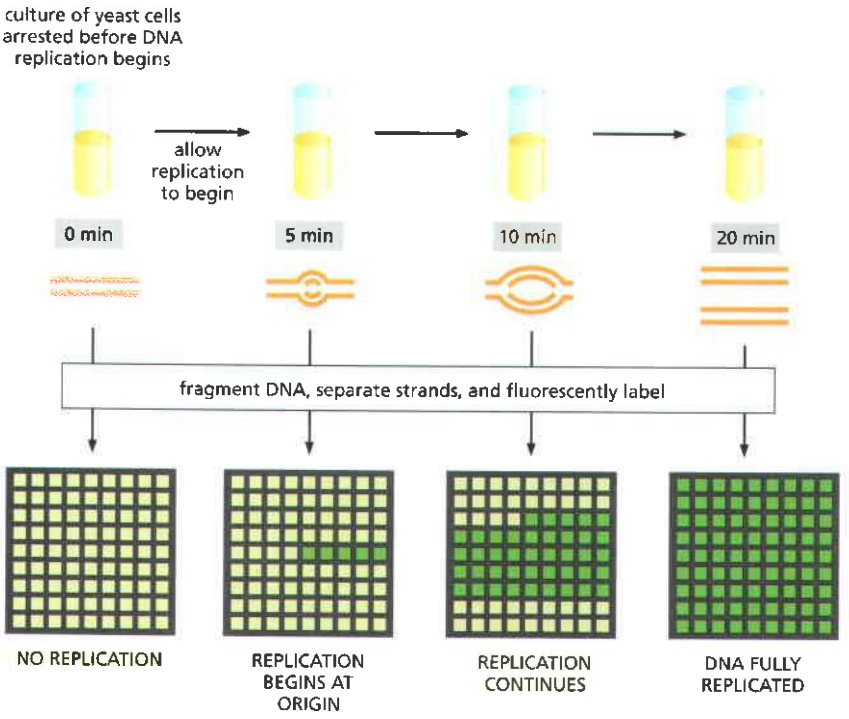


Figure 5–32 Use of DNA microarrays to monitor the formation and progress of replication forks in the budding yeast genome. For this experiment, a population of cells is sychronized so that they all begin replication at the same time. DNA is collected and hybridized to the microarray; DNA that has been replicated once gives a hybridization signal *(dark green squares)* twice as high as that of unreplicated DNA *(light green squares)*. The spots on these microarrays represent consecutive sequences along a segment of a yeast chromosome arranged left to right, top to bottom. Only 81 spots are shown here, but the actual arrays contain tens of thousands of sequences that span the entire yeast genome. As can be seen, replication begins at an origin and proceeds bidirectionally. For simplicity only one origin is shown here. In yeast cells, replication begins at hundreds of origins located throughout the genome.

These findings suggest that regions of the genome whose chromatin is least condensed are replicated first. Replication forks seem to move at comparable rates throughout S phase, so the extent of chromosome condensation seems to influence the time at which replication forks are initiated, rather than their speed once formed.

Well-defined DNA Sequences Serve as Replication Origins in a Simple Eucaryote, the Budding Yeast

Having seen that a eucaryotic chromosome is replicated using many origins of replication, each of which "fires" at a characteristic time in S phase of the cell cycle, we turn to the nature of these origins of replication. We saw earlier in this chapter that replication origins have been precisely defined in bacteria as specific DNA sequences that attract initiator proteins, which then assemble the DNA replication machinery. By analogy, one would expect the replication origins in eucaryotic chromosomes to be specific DNA sequences too.

The search for eucaryotic DNA sequences that carry all the information necessary to specify a replication origin has been most productive in the budding yeast *S. cerevisiae*. Powerful selection methods to find them have been devised that make use of mutant yeast cells defective for an essential gene. These cells can survive in a selective medium only if they are provided with DNA that carries a functional copy of the missing gene. If a circular bacterial plasmid containing this gene is introduced into the mutant yeast cells directly, the plasmid will not be able to replicate because it lacks a functional origin. If random pieces of yeast DNA are inserted into this plasmid, however, only the rare plasmid DNA molecules that happen to contain a yeast replication origin can replicate. The yeast cells that carry such plasmids are able to proliferate because they have been provided with the essential gene in a form that can be replicated and passed on to progeny cells (**Figure 5–33**). A DNA sequence identified by its presence in a plasmid isolated from these surviving yeast cells is called an *autonomously replicating sequence (ARS)*. Most ARSs have been shown to be authentic chromosomal origins of replication, thereby validating the strategy used to obtain them.

For budding yeast, the location of every origin of replication on each chromosome has been determined. The particular chromosome shown in **Figure 5–34**—chromosome III from the yeast *S. cerevisiae*—is one of the smallest

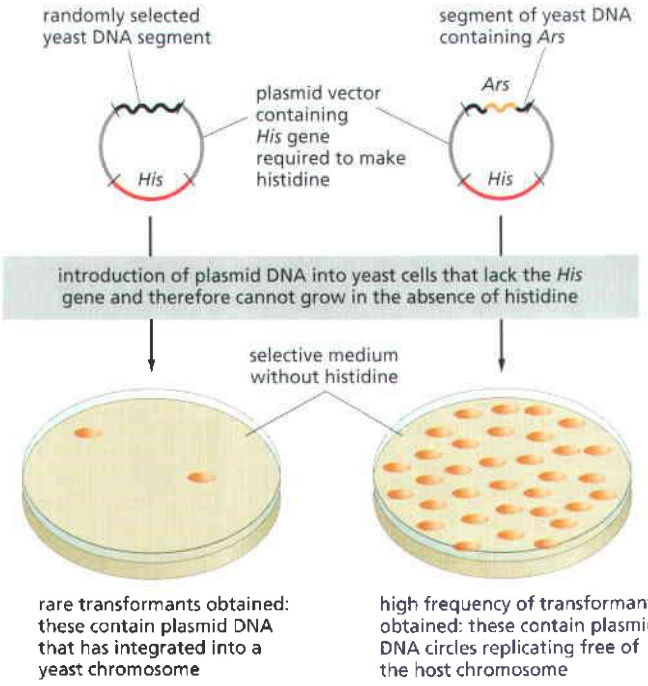


Figure 5–33 A strategy used to identify DNA sequences that are sufficient for initiating DNA replication. Each of the yeast DNA sequences identified in this way was called an autonomously replicating sequence *(Ars)*, since it enables a plasmid that contains it to replicate in the host cell without having to be incorporated into a host cell chromosome.

chromosomes known, with a length less than 1/100 that of a typical human chromosome. Its major origins are spaced an average of 30,000 nucleotide pairs apart; this density of origins should permit this chromosome to be replicated in about 10 minutes if all the origins fire at once. As we have previously discussed, mammalian origins are spaced further apart, typically every 100,000 to 250,000 nucleotide pairs.

Genetic experiments in *S. cerevisiae* have tested the effect of deleting various replication origins on chromosome III. Removing a few origins has little effect, because replication forks that begin at neighboring origins of replication can continue into the regions that lack their own origins. The deletion of more replication origins, however, results in the loss of the chromosome as the cells divide, because it is replicated too slowly. Many eucaryotes carry an excess of origins of replication, presumably to ensure that the complete genome can still be replicated in a timely manner if some of the origins fail to function.

A Large Multisubunit Complex Binds to Eucaryotic Origins of Replication

The minimal DNA sequence required for directing DNA replication initiation in the yeast *S. cerevisiae* has been determined by testing smaller and smaller DNA fragments in the experiment shown in Figure 5–33. Each DNA sequence that can serve as an origin of replication is found to contain (1) a binding site for a large, multisubunit initiator protein called **ORC**, for **origin recognition complex**, (2) a stretch of DNA that is rich in As and Ts and therefore easy to unwind, and (3) at least one binding site for proteins that help attract ORC to the origin DNA (**Figure 5–35**). In bacteria, once the initiator protein is properly bound to the single origin of replication the assembly of the replication forks follows more or less automatically. In eucaryotes, the situation is significantly different because of a profound problem eucaryotes have in replicating chromosomes with so many origins of replication (an estimated 400 in yeast and 10,000 in

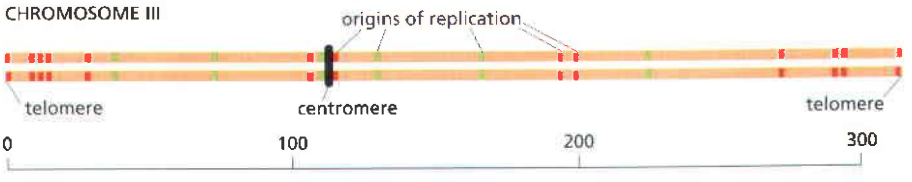


Figure 5–34 The origins of DNA replication on chromosome III of the yeast *S. cerevisiae*. This chromosome, one of the smallest eucaryotic chromosomes known, carries a total of 180 genes. As indicated, it contains 19 replication origins, although they are used with different efficiencies. Those in *red* are typically used less than 10% of the time, while those in *green* are used in 90% of S phases.



Figure 5–35 An origin of replication in yeast. Comprising about 150 nucleotide pairs, this yeast origin (identified by the procedure shown in Figure 5–33) has a binding site for ORC, and one for Abf1, an auxillary protein that facilitates ORC binding. All origins contain binding sites for ORC but the auxiliary proteins are different from one origin to the next. Most origins, like the one depicted, also contain a stretch of DNA that is especially easy to unwind.

humans, for example). With so many places to begin replication, how is the process regulated to ensure that all the DNA is copied once and only once?

The answer lies in the way that the ORC complex, once bound to an origin of replication, is sequentially activated and deactivated. This matter is discussed in detail in Chapter 17, when we consider the cellular machinery that underlies the cell division cycle. The ORC–origin interaction persists throughout the entire cell cycle, dissociating only briefly immediately following replication of the origin DNA, and other proteins that bind to it regulate origin activity. These include the DNA helicase and two helicase loading proteins, Cdc6 and Cdt1, which are assembled onto an ORC–DNA complex to form a *prereplicative complex* at each origin during G_1 phase (**Figure 5–36**). The passage of a cell from G_1 to S phase is triggered by the activation of protein kinases (Cdks) that lead to dissociation of the helicase loading proteins, activation of the helicase, unwinding of the origin DNA, and loading of the remaining replication proteins including DNA polymerases (see Figure 5–36).

The protein kinases that trigger DNA replication simultaneously prevent all assembly of new prereplicative complexes until the next M phase resets the entire cycle (for details, see pp. 1067–1069). This strategy provides a single window of opportunity for prereplicative complexes to form (G_1 phase, when Cdk activity is low) and a second window for them to be activated and subsequently disassembled (S phase, when Cdk activity is high). Because these two phases of the cell cycle are mutually exclusive and occur in a prescribed order, each origin of replication can fire once and only once a cell cycle.

The Mammalian DNA Sequences That Specify the Initiation of Replication Have Been Difficult to Identify

Compared with the situation in budding yeasts, DNA sequences that specify replication origins in other eucaryotes have been more difficult to define. Recently, however, it has been possible to identify specific human DNA sequences, each several thousand nucleotide pairs in length, that are sufficient to serve as replication origins. These origins continue to function when moved to a different chromosomal region by recombinant DNA methods, as long as they are placed in a region where the chromatin is relatively uncondensed. One of these origins is from the β-globin gene cluster. At its normal position in the genome, the function of this origin depends critically upon distant DNA sequences (**Figure 5–37**). As discussed in Chapter 7, this distant DNA is needed for expression of all genes in the β-globin cluster, and its effects on both transcription and origin function may reflect its long-range decondensation of chromatin structure.

A human ORC homologous to that in yeast cells is required for replication initiation. Many of the other proteins that function in the initiation process in yeast likewise have central roles in humans. It therefore seems likely that the yeast and human initiation mechanisms will turn out to be similar in outline. However, the binding sites for the ORC protein seem to be less specific in humans than they are in yeast, which may explain why the replication origins of humans are less sharply defined. In fact, chromatin structure, rather than DNA sequences, may have the central role in defining mammalian origins of replication. Thus, as is true in many other areas of cell biology, the mechanism of DNA replication initiation in yeast may vividly highlight the core process, while the situation in humans represents an elaborate variation on the theme.

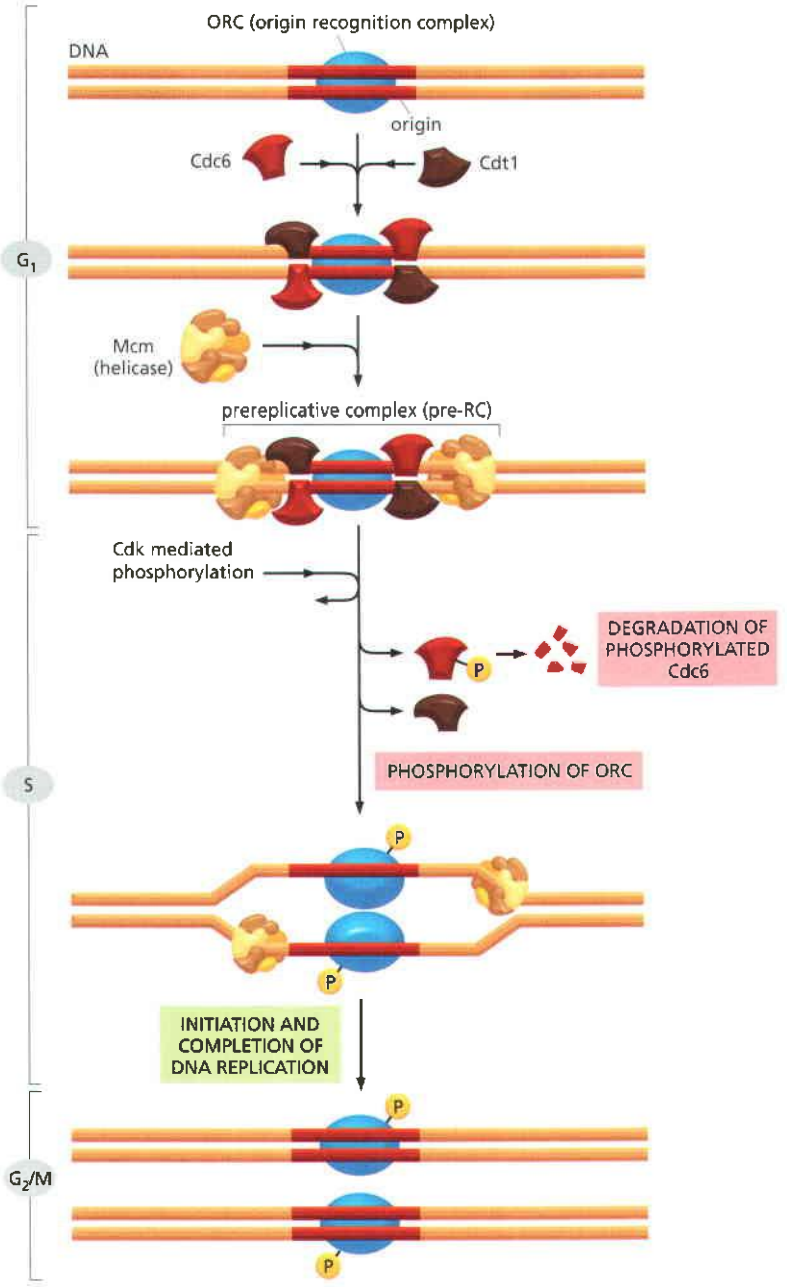


Figure 5–36 The mechanism of DNA replication initiation in eucaryotes. This mechanism ensures that each origin of replication is activated only once per cell cycle. An origin of replication can be used only if a prereplicative complex forms there in G_1 phase. At the beginning of S phase, cyclin-dependent kinases (Cdks) phosphorylate various replication proteins, causing both disassembly of the prereplicative complex and initiation of DNA replication. A new prereplicative complex cannot form at the origin until the cell progresses to the next G_1 phase.

New Nucleosomes Are Assembled Behind the Replication Fork

There are several additional aspects of DNA replication that are specific to eucaryotes. As discussed in Chapter 4, eucaryotic chromosomes are composed of roughly equal mixtures of DNA and protein. Chromosome duplication therefore requires not only the replication of DNA, but also the synthesis and assembly of new chromosomal proteins onto the DNA behind each replication fork. Although we are far from understanding this process in detail, we are beginning to learn how the fundamental unit of chromatin packaging, the nucleosome, is duplicated. The cell requires a large amount of new histone protein, approximately equal in mass to the newly synthesized DNA, to make the new nucleosomes in each cell cycle. For this reason, most eucaryotic organisms possess multiple copies of the gene for each histone. Vertebrate cells, for example, have about 20 repeated gene sets, most containing the genes that encode all five histones (H1, H2A, H2B, H3, and H4).

Unlike most proteins, which are made continuously throughout interphase, histones are synthesized mainly in S phase, when the level of histone mRNA increases about fiftyfold as a result of both increased transcription and

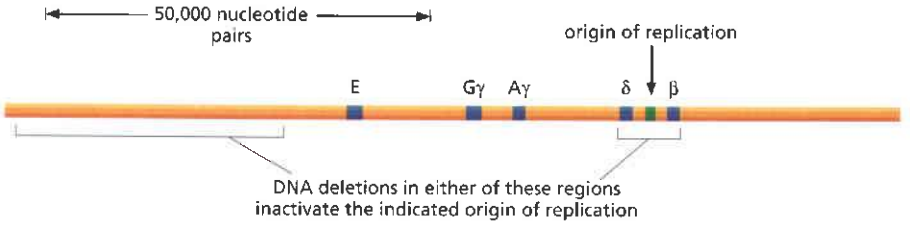


Figure 5–37 Deletions that inactivate an origin of replication in humans. These two deletions are found separately in two individuals who suffer from *thalassemia*, a disorder caused by the failure to express one or more of the genes in the β-globin gene cluster shown. In both of these deletion mutants, the DNA in this region is replicated by forks that begin at replication origins outside the β-globin gene cluster.

decreased mRNA degradation. The major histone mRNAs are degraded within minutes when DNA synthesis stops at the end of S phase. The mechanism depends on special properties of the 3′ ends of these mRNAs, as discussed in Chapter 7. In contrast, the histone proteins themselves are remarkably stable and may survive for the entire life of a cell. The tight linkage between DNA synthesis and histone synthesis appears to reflect a feedback mechanism that monitors the level of free histone to ensure that the amount of histone made exactly matches the amount of new DNA synthesized.

As a replication fork advances, it must somehow pass through the parental nucleosomes. *In vitro* studies show that the replication apparatus has a poorly understood intrinsic ability to pass through parental nucleosomes without displacing them from the DNA. However, to replicate chromosomes efficiently in the cell, chromatin-remodeling proteins (discussed in Chapter 4), which destabilize the DNA–histone interface, are required. Aided by such complexes, replication forks can transit even highly condensed heterochromatin efficiently.

As the replication fork passes through chromatin, most of the old histones remain DNA-bound and are distributed to the daughter DNA helices behind a replication fork. But since the amount of DNA has doubled, an equal amount of new histones is also needed to complete the packaging of DNA into chromatin. The old and the new histones are combined in an intriguing way. When a nucleosome is transited by a replication fork, the histone octamer appears to be broken into an H3-H4 tetramer and two H2A-H2B dimers (see Figure 4–26). The H3-H4 tetramer remains associated with DNA and is distributed at random to one or the other daughter duplexes, but the H2A-H2B dimers are released from DNA. Freshly made H3-H4 tetramers are added to the newly synthesized DNA to fill in the "spaces," and H2AB dimers—half of which are old and half new—are then added at random to complete the nucleosomes (**Figure 5–38**).

The orderly and rapid addition of new H3-H4 tetramers and H2A-H2B dimers behind a replication fork requires **histone chaperones** (also called *chromatin assembly factors*). These multisubunit complexes bind the highly basic histones and release them for assembly only in the appropriate context. These histone chaperones, along with their cargoes, are directed to newly replicated DNA through a specific interaction with the eucaryotic sliding clamp, called *PCNA* (see Figure 5–38B). These clamps are left behind moving replication forks and remain on the DNA long enough for the histone chaperones to complete their tasks.

The Mechanisms of Eucaryotic Chromosome Duplication Ensure That Patterns of Histone Modification Can Be Inherited

We saw in Chapter 4 that histones are subject to many types of covalent modifications and that the patterns of these modifications can carry important information regarding the fate of underlying DNA. It makes little intuitive sense for these patterns to be erased each time a cell divides, but since this information is encoded in the histone proteins rather than the DNA, special mechansims are needed to preserve and duplicate it. We have seen that histone H3-H4 tetramers are distributed randomly to the two daughter chromosomes that emerge behind a moving replication fork. The tails as well as other regions of H3 and H4 can be extensively modified (see Figure 4–39), and thus each daughter chromosome is seeded with the memory of the parental pattern of H3 and H4 modification.

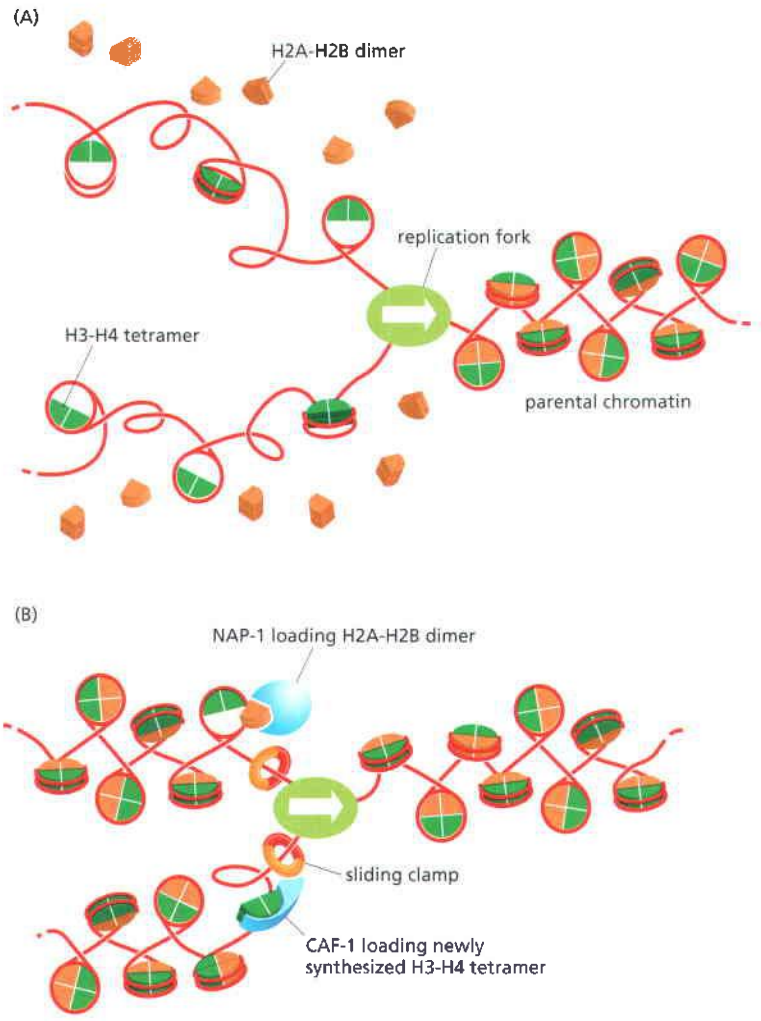


Figure 5–38 Distribution of parental and newly synthesized histones behind a eucaryotic replication fork. (A) The distribution of parental H3-H4 tetramers to the daughter DNA molecules is apparently random, with roughly equal numbers inherited by each daughter. In contrast, H2A-H2B dimers are released from the DNA as the replication fork passes. (B) Histone chaperones (NAP1 and CAF1) restore the full complement of histones to daughter molecules. Although some daughter nucleosomes contain only parental histones or only newly synthesized histones, most are hybrids of old and new. (Adapted from J.D. Watson et al., Molecular Biology of the Gene, 5th ed. Cold Spring Harbor: Cold Spring Harbor Laboratory Press, 2004.)

Once the nucleosome assembly behind a replication fork has been completed, the parental patterns of H3-H4 modification can be reinforced through histone modification enzymes in reader–writer complexes that recognize the same type of modification they create (**Figure 5–39**).

The faithful duplication of patterns of histone modification may be responsible for many examples of *epigenetic inheritance*, in which a heritable change in a cell's phenotype occurs without a change in the nucleotide sequence of DNA. We shall revisit the topic of epigenetics in Chapter 7 when we consider how decisions made by a cell are "remembered" by its progeny cells many generations later.

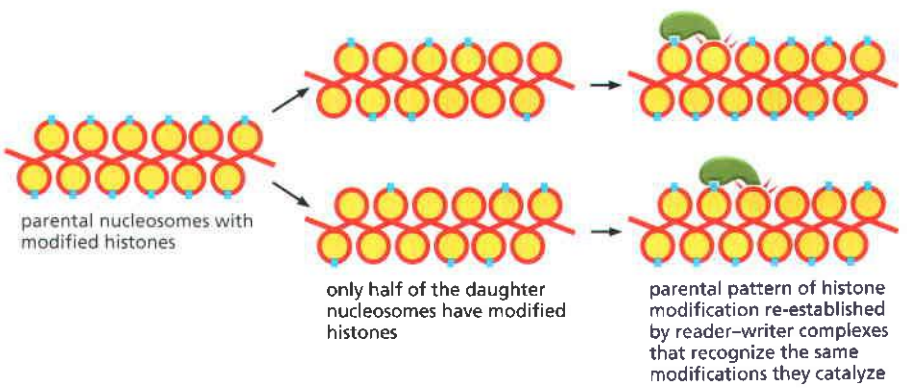


Figure 5–39 Strategy through which parental patterns of histone H3 and H4 modification can be inherited by daughter chromosomes. Although it is unlikely that this mechanism applies to all histone modifications, it does pertain to some (see Figure 4–51). For example, a number of histone methylase complexes specifically recognize N-terminal histone tails that have been previously methylated at the same site that the methylase modifies.

Telomerase Replicates the Ends of Chromosomes

We saw earlier that synthesis of the lagging strand at a replication fork must occur discontinuously through a backstitching mechanism that produces short DNA fragments. This mechanism encounters a special problem when the replication fork reaches an end of a linear chromosome: there is no place to produce the RNA primer needed to start the last Okazaki fragment at the very tip of a linear DNA molecule.

Bacteria solve this "end-replication" problem by having circular DNA molecules as chromosomes (see Figure 5–27). Eucaryotes solve it in an ingenious way: they have specialized nucleotide sequences at the ends of their chromosomes that are incorporated into structures called telomeres (see Chapter 4). Telomeres contain many tandem repeats of a short sequence that is similar in organisms as diverse as protozoa, fungi, plants, and mammals. In humans, the sequence of the repeat unit is GGGTTA, and it is repeated roughly a thousand times at each telomere.

Telomere DNA sequences are recognized by sequence-specific DNA-binding proteins that attract an enzyme, called **telomerase**, that replenishes these sequences each time a cell divides. Telomerase recognizes the tip of an existing telomere DNA repeat sequence and elongates it in the 5′-to-3′ direction, using an RNA template that is a component of the enzyme itself to synthesize new copies of the repeat (**Figure 5–40**). The enzymatic portion of telomerase resembles other *reverse transcriptases*, enzymes that synthesize DNA using an RNA template (see Figure 5–72). After extension of the parental DNA strand by telomerase, replication of the lagging strand at the chromosome end can be completed by the conventional DNA polymerases, using these extensions as a template to synthesize the complementary strand (**Figure 5–41**).

The mechanism just described, aided by a nuclease that chews back the 5′ end, ensures that the 3′ DNA end at each telomere is always longer than the 5′ end with which it is paired, leaving a protruding single-stranded end (see Figure 5–41). This protruding end has been shown to loop back and tuck its single-stranded terminus into the duplex DNA of the telomeric repeat sequence to form a *t-loop* (**Figure 5–42**). In broad outline this reaction resembles strand invasion during homologous recombination, discussed below, and it may have evolved from these ancient recombination systems. T-loops provide the normal ends of chromosomes with a unique structure, which protects them from degradative enzymes and clearly distinguishes them from the ends of the broken DNA molecules that the cell rapidly repairs (see Figure 5–51).

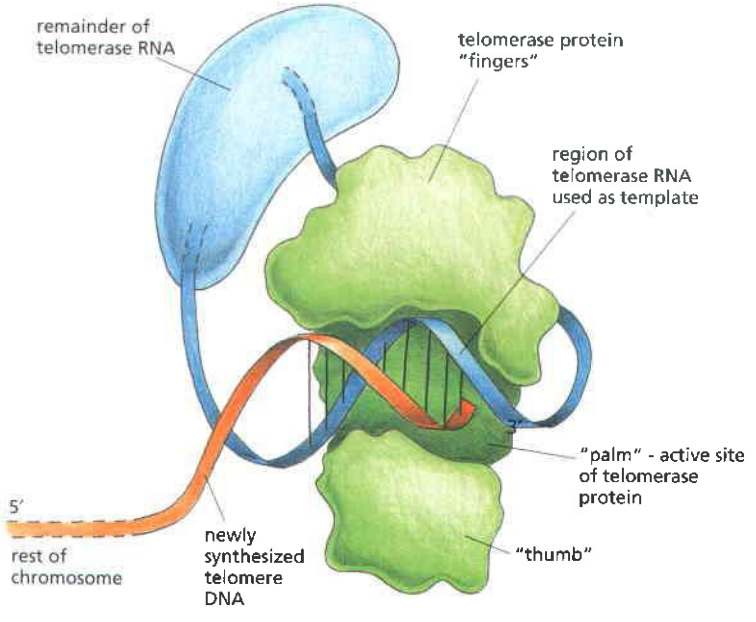


Figure 5–40 The structure of a portion of telomerase. Telomerase is a large protein–RNA complex. The RNA *(blue)* contains a templating sequence for synthesizing new DNA telomere repeats. The synthesis reaction itself is carried out by the reverse transcriptase domain of the protein, shown in *green*. A reverse transcriptase is a special form of polymerase enzyme that uses an RNA template to make a DNA strand; telomerase is unique in carrying its own RNA template with it at all times. Telomerase also has several additional protein domains (not shown) that are needed to assemble the enzyme at the ends of chromosomes properly. (Modified from J. Lingner and T.R. Cech, *Curr. Opin. Genet. Dev.* 8:226–232, 1998. With permission from Elsevier.)

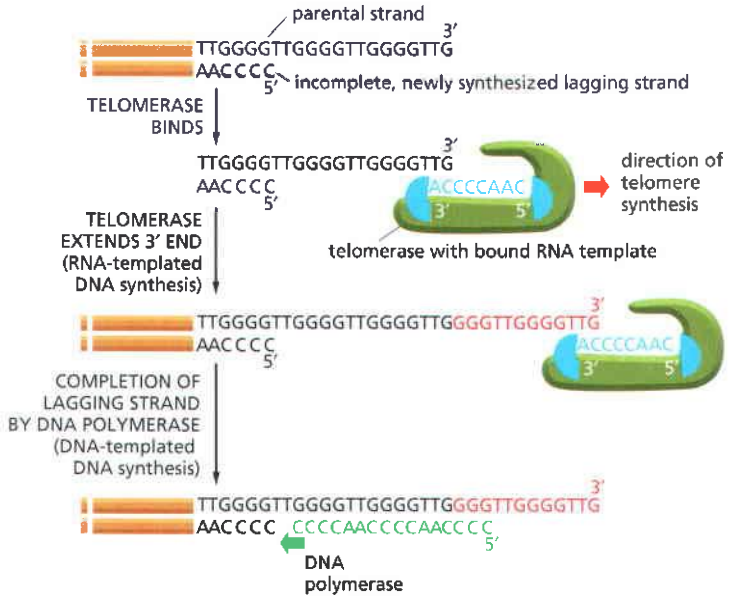


Figure 5–41 Telomere replication. Shown here are the reactions that synthesize the repeating G-rich sequences that form the ends of the chromosomes (telomeres) of diverse eucaryotic organisms. The 3′ end of the parental DNA strand is extended by RNA-templated DNA synthesis; this allows the incomplete daughter DNA strand that is paired with it to be extended in its 5′ direction. This incomplete, lagging strand is presumed to be completed by DNA polymerase α, which carries a DNA primase as one of its subunits. The telomere sequence illustrated is that of the ciliate *Tetrahymena*, in which these reactions were first discovered.

Telomere Length Is Regulated by Cells and Organisms

Because the processes that grow and shrink each telomere sequence are only approximately balanced, a chromosome end contains a variable number of telomeric repeats. Not surprisingly, experiments show that cells that proliferate indefinitely (such as yeast cells) have homeostatic mechanisms that maintain the number of these repeats within a limited range (**Figure 5–43**).

In the somatic cells of humans, the telomere repeats have been proposed to provide each cell with a counting mechanism that helps prevent the unlimited proliferation of wayward cells in adult tissues. The simplest form of this idea holds that our somatic cells are born with a full complement of telomeric repeats. Some stem cells, notably those in tissues that must be replenished throughout life—such as those of bone marrow or skin—retain full telomerase activity. However, in many other types of cells, the level of telomerase is turned down so that the enzyme cannot quite keep up with chromosome duplication. Such cells lose 100–200 nucleotides from each telomere every time they divide. After many cell generations, the descendent cells will inherit defective chromosomes (because their tips cannot be replicated completely) and consequently will withdraw permanently from the cell cycle and cease dividing—a process called *replicative cell senescence* (discussed in Chapter 17). In theory, such a mechanism could provide a safeguard against the uncontrolled cell proliferation of abnormal cells in somatic tissues, thereby helping to protect us from cancer.

The idea that telomere length acts as a "measuring stick" to count cell divisions and thereby regulate the cell's lifetime has been tested in several ways. For certain types of human cells grown in tissue culture, the experimental results support such a theory. Human fibroblasts normally proliferate for about 60 cell divisions in culture before undergoing replicative senescence. Like most other somatic cells in humans, fibroblasts produce only low levels of telomerase, and their telomeres gradually shorten each time they divide. When telomerase is provided to the fibroblasts by inserting an active telomerase gene, telomere length is maintained and many of the cells now continue to proliferate indefinitely. It therefore seems clear that telomere shortening can count cell divisions and trigger replicative senescence in some human cells.

It has been proposed that this type of control on cell proliferation is important for the maintenance of tissue architecture and that it is also somehow responsible for the aging of animals like ourselves. These ideas have been tested by producing transgenic mice that lack telomerase entirely. The telomeres in mouse chromosomes are about five times longer than human telomeres, and the mice must therefore be bred through three or more generations before their telomeres have shrunk to the normal human length. It is therefore perhaps not

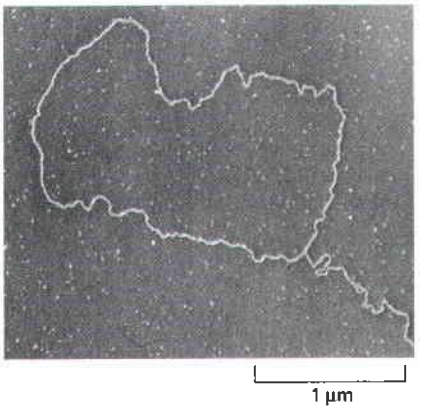


Figure 5–42 A t-loop at the end of a mammalian chromosome. Electron micrograph of the DNA at the end of an interphase human chromosome. The chromosome was fixed, deproteinated, and artificially thickened before viewing. The loop seen here is approximately 15,000 nucleotide pairs in length. The insertion of the single-stranded 3′ end into the duplex repeats to form a t-loop is thought to be carried out and maintained by specialized proteins. (From J.D. Griffith et al., *Cell* 97:503–514, 1999. With permission from Elsevier.)

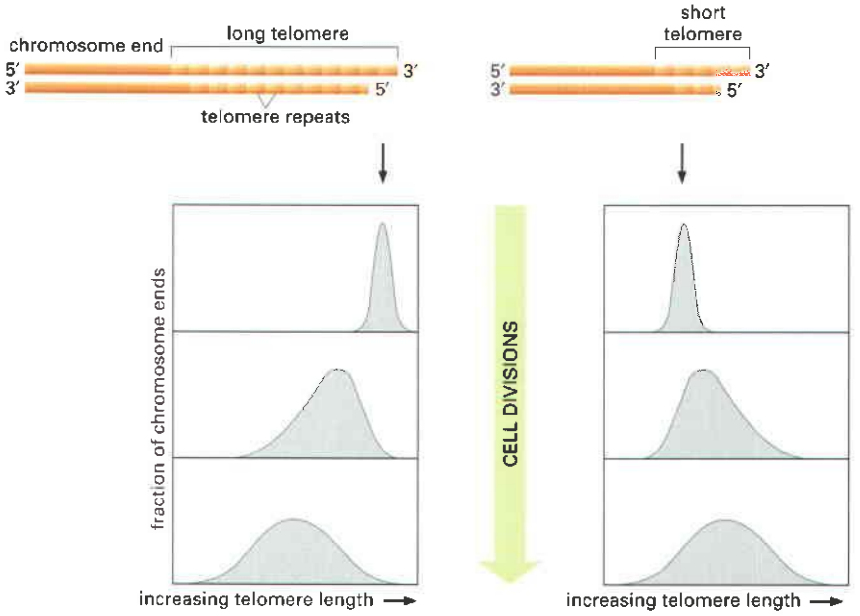


Figure 5–43 A demonstration that yeast cells control the length of their telomeres. In this experiment, the telomere at one end of a particular chromosome is artificially made either longer *(left)* or shorter *(right)* than average. After many cell divisions, the chromosome recovers, showing an average telomere length and a length distribution that is typical of the other chromosomes in the yeast cell. A similar feedback mechanism for controlling telomere length has been proposed for the cells in the germ-line cells of animals.

surprising that the first generations of mice develop normally. However, the mice in later generations develop progressively more defects in some of their highly proliferative tissues. In addition, these mice show signs of premature aging and have a pronounced tendency to develop tumors. In these and other respects these mice resemble humans with the genetic disease *dyskeratosis congenita*. Individuals afflicted with this disease carry one functional and one nonfunctional copy of the telomerase RNA gene; they have prematurely shortened telomeres and typically die of progressive bone marrow failure. They also develop lung scarring and liver cirrhosis and show abnormalities in various epidermal structures including skin, hair follicles, and nails.

The above observations demonstrate that controlling cell proliferation by telomere shortening poses a risk to an organism, because not all of the cells that begin losing the ends of their chromosomes will stop dividing. Some apparently become genetically unstable, but continue to divide giving rise to variant cells that can lead to cancer. Thus, one can question whether the observed down-regulation of telomerase in most human somatic cells provides an evolutionary advantage, as suggested by those who postulate that telomere shortening protects us from cancer and other proliferative diseases.

Summary

The proteins that initiate DNA replication bind to DNA sequences at a replication origin to catalyze the formation of a replication bubble with two outward-moving replication forks. The process begins when an initiator protein–DNA complex is formed that subsequently loads a DNA helicase onto the DNA template. Other proteins are then added to form the multienzyme "replication machine" that catalyzes DNA synthesis at each replication fork.

In bacteria and some simple eucaryotes, replication origins are specified by specific DNA sequences that are only several hundred nucleotide pairs long. In other eucaryotes, such as humans, the sequences needed to specify an origin of DNA replication seem to be less well defined, and the origin can span several thousand nucleotide pairs.

Bacteria typically have a single origin of replication in a circular chromosome. With fork speeds of up to 1000 nucleotides per second, they can replicate their genome in less than an hour. Eucaryotic DNA replication takes place in only one part of the cell cycle, the S phase. The replication fork in eucaryotes moves about 10 times more slowly than the bacterial replication fork, and the much longer eucaryotic chromosomes each require many replication origins to complete their replication in an S phase, which typically lasts for 8 hours in human cells. The different replication origins in these

eucaryotic chromosomes are activated in a sequence, determined in part by the structure of the chromatin, with the most condensed regions of chromatin typically beginning their replication last. After the replication fork has passed, chromatin structure is re-formed by the addition of new histones to the old histones that are directly inherited by each daughter DNA molecule. The mechanism of chromosome duplication allows the parental patterns of histone modification to be passed on to daughter chromosomes, thus providing a means of epigenetic inheritance.

Eucaryotes solve the problem of replicating the ends of their linear chromosomes with a specialized end structure, the telomere, maintained by a special nucleotide polymerizing enzyme called telomerase. Telomerase extends one of the DNA strands at the end of a chromosome by using an RNA template that is an integral part of the enzyme itself, producing a highly repeated DNA sequence that typically extends for thousands of nucleotide pairs at each chromosome end.

DNA REPAIR

Maintaining the genetic stability that an organism needs for its survival requires not only an extremely accurate mechanism for replicating DNA, but also mechanisms for repairing the many accidental lesions that occur continually in DNA. Most such spontaneous changes in DNA are temporary because they are immediately corrected by a set of processes that are collectively called **DNA repair**. Of the thousands of random changes created every day in the DNA of a human cell by heat, metabolic accidents, radiation of various sorts, and exposure to substances in the environment, only a few accumulate as mutations in the DNA sequence. For example, we now know that fewer than one in 1000 accidental base changes in DNA results in a permanent mutation; the rest are eliminated with remarkable efficiency by DNA repair.

The importance of DNA repair is evident from the large investment that cells make in DNA repair enzymes. For example, analysis of the genomes of bacteria and yeasts has revealed that several percent of the coding capacity of these organisms is devoted solely to DNA repair functions. The importance of DNA repair is also demonstrated by the increased rate of mutation that follows the inactivation of a DNA repair gene. Many DNA repair proteins and the genes that encode them—which we now know operate in a wide range of organisms, including humans—were originally identified in bacteria by the isolation and characterization of mutants that displayed an increased mutation rate or an increased sensitivity to DNA-damaging agents.

Recent studies of the consequences of a diminished capacity for DNA repair in humans have linked many human diseases with decreased repair (Table 5–2).

Table 5–2 Some Inherited Syndromes with Defects in DNA Repair

NAME	PHENOTYPE	ENZYME OR PROCESS AFFECTED
MSH2, 3, 6, MLH1, PMS2	colon cancer	mismatch repair
Xeroderma pigmentosum (XP) groups A–G	skin cancer, UV sensitivity, neurological abnormalities	nucleotide excision–repair
XP variant	UV sensitivity, skin cancer	translesion synthesis by DNA polymerase η
Ataxia telangiectasia (AT)	leukemia, lymphoma, γ-ray sensitivity, genome instability	ATM protein, a protein kinase activated by double-strand breaks
BRCA2	breast, ovarian, and prostate cancer	repair by homologous recombination
Werner syndrome	premature aging, cancer at several sites, genome instability	accessory 3′-exonuclease and DNA helicase
Bloom syndrome	cancer at several sites, stunted growth, genome instability	accessory DNA helicase for replication
Fanconi anemia groups A–G	congenital abnormalities, leukemia, genome instability	DNA interstrand cross-link repair
46 BR patient	hypersensitivity to DNA-damaging agents, genome instability	DNA ligase I

Thus, we saw previously that defects in a human gene that normally functions to repair the mismatched base pairs resulting from DNA replication errors can lead to an inherited predisposition to certain cancers, reflecting an increased mutation rate. In another human disease, *xeroderma pigmentosum* (XP), the afflicted individuals have an extreme sensitivity to ultraviolet radiation because they are unable to repair certain DNA photoproducts. This repair defect results in an increased mutation rate that leads to serious skin lesions and an increased susceptibility to certain cancers.

Without DNA Repair, Spontaneous DNA Damage Would Rapidly Change DNA Sequences

Although DNA is a highly stable material, as required for the storage of genetic information, it is a complex organic molecule that is susceptible, even under normal cell conditions, to spontaneous changes that would lead to mutations if left unrepaired (**Figure 5–44**). For example, the DNA of each human cell loses about 5000 purine bases (adenine and guanine) every day because their N-glycosyl linkages to deoxyribose hydrolyze, a spontaneous reaction called *depurination.* Similarly, a spontaneous *deamination* of cytosine to uracil in DNA occurs at a rate of about 100 bases per cell per day (**Figure 5–45**). DNA bases are also occasionally damaged by an encounter with reactive metabolites produced in the cell (including reactive forms of oxygen) or by exposure to chemicals in the environment. Likewise, ultraviolet radiation from the sun can produce a covalent linkage between two adjacent pyrimidine bases in DNA to form, for example, thymine dimers (**Figure 5–46**). If left uncorrected when the DNA is replicated, most of these changes would be expected to lead either to the deletion of one or more base pairs or to a base-pair substitution in the daughter DNA chain (**Figure 5–47**). The mutations would then be propagated throughout subsequent cell generations. Such a high rate of random changes in the DNA sequence would have disastrous consequences for an organism.

The DNA Double Helix Is Readily Repaired

The double-helical structure of DNA is ideally suited for repair because it carries two separate copies of all the genetic information—one in each of its two strands. Thus, when one strand is damaged, the complementary strand retains an intact copy of the same information, and this copy is generally used to restore the correct nucleotide sequences to the damaged strand.

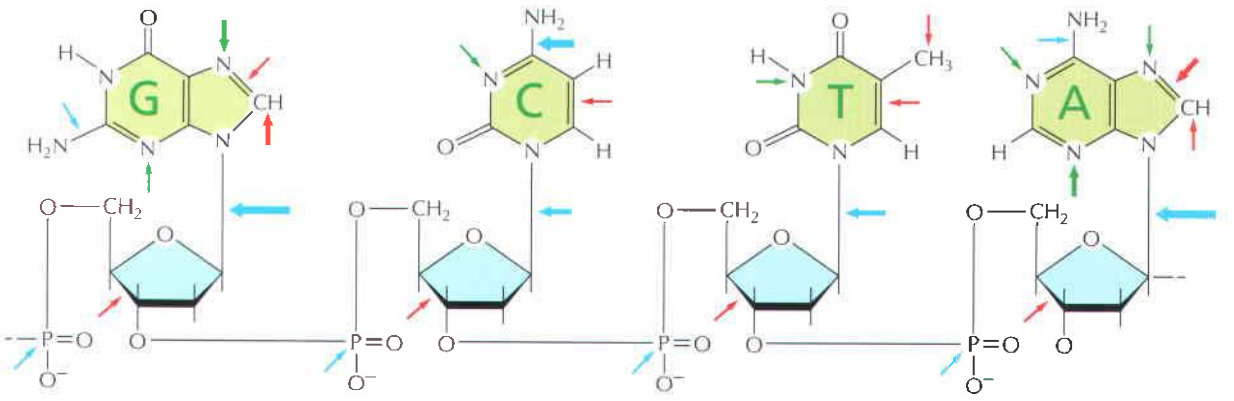


Figure 5–44 A summary of spontaneous alterations likely to require DNA repair. The sites on each nucleotide that are known to be modified by spontaneous oxidative damage *(red arrows),* hydrolytic attack *(blue arrows),* and uncontrolled methylation by the methyl group donor *S*-adenosylmethionine *(green arrows)* are shown, with the width of each arrow indicating the relative frequency of each event. (After T. Lindahl, *Nature* 362:709–715, 1993. With permission from Macmillan Publishers Ltd.)

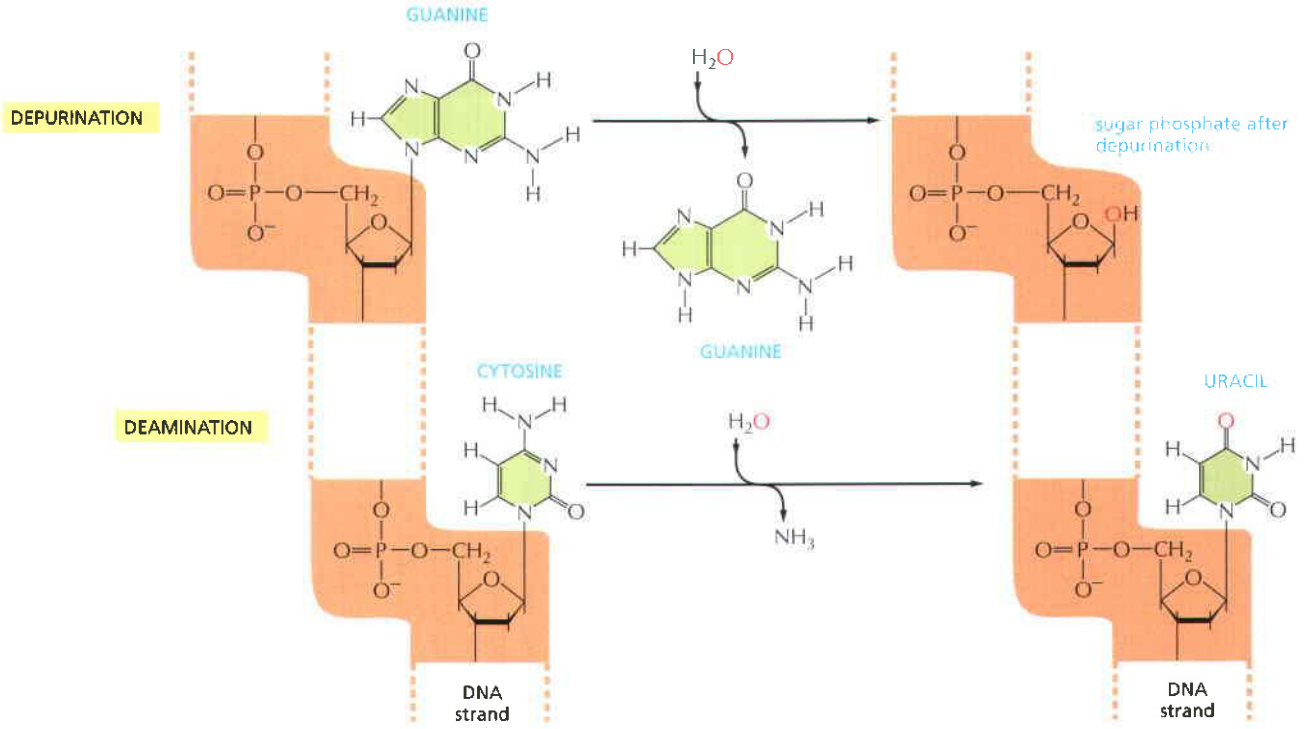


Figure 5–45 Depurination and deamination. These two reactions are the most frequent spontaneous chemical reactions known to create serious DNA damage in cells. Depurination can release guanine (shown here), as well as adenine, from DNA. The major type of deamination reaction converts cytosine to an altered DNA base, uracil (shown here), but deamination occurs on other bases as well. These reactions normally take place in double-helical DNA; for convenience, only one strand is shown.

An indication of the importance of a double-stranded helix to the safe storage of genetic information is that all cells use it; only a few small viruses use single-stranded DNA or RNA as their genetic material. The types of repair processes described in this section cannot operate on such nucleic acids, and once damaged, the chance of a permanent nucleotide change occurring in these single-stranded genomes of viruses is thus very high. It seems that only organisms with tiny genomes (and therefore tiny targets for DNA damage) can afford to encode their genetic information in any molecule other than a DNA double helix.

DNA Damage Can Be Removed by More Than One Pathway

Cells have multiple pathways to repair their DNA using different enzymes that act upon different kinds of lesions. **Figure 5–48** shows two of the most common pathways. In both, the damage is excised, the original DNA sequence is restored by a DNA polymerase that uses the undamaged strand as its template, and a remaining break in the double helix is sealed by DNA ligase (see Figure 5–13).

The two pathways differ in the way in which they remove the damage from DNA. The first pathway, called **base excision repair**, involves a battery of enzymes called *DNA glycosylases*, each of which can recognize a specific type of altered base in DNA and catalyze its hydrolytic removal. There are at least six types of these enzymes, including those that remove deaminated Cs, deaminated As, different types of alkylated or oxidized bases, bases with opened rings, and bases in which a carbon–carbon double bond has been accidentally converted to a carbon–carbon single bond. How is an altered base detected within the context of the double helix? A key step is an enzyme-mediated "flipping-out"

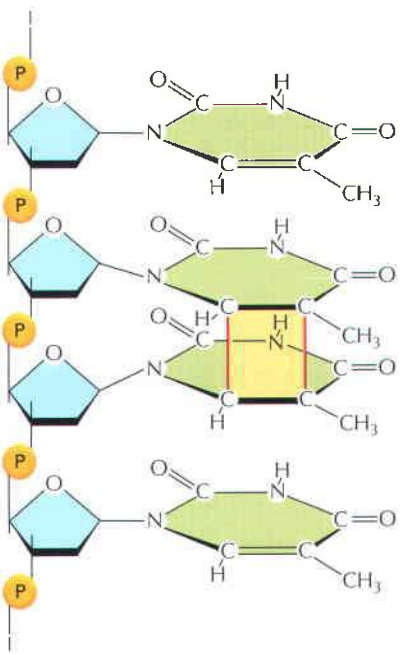


Figure 5–46 The most common type of thymine dimer. This type of damage occurs in the DNA of cells exposed to ultraviolet irradiation (as in sunlight). A similar dimer will form between any two neighboring pyrimidine bases (C or T residues) in DNA.

of the altered nucleotide from the helix, which allows the DNA glycosylase to probe all faces of the base for damage (**Figure 5–49**). It is thought that these enzymes travel along DNA using base-flipping to evaluate the status of each base. Once an enzyme finds the damaged base that it recognizes, it removes the base from its sugar.

The "missing tooth" created by DNA glycosylase action is recognized by an enzyme called *AP endonuclease* (AP for *apurinic* or *apyrimidinic, endo* to signify that the nuclease cleaves within the polynucleotide chain), which cuts the phosphodiester backbone, after which the damage is removed and the resulting gap repaired (see Figure 5–48A). Depurination, which is by far the most frequent type of damage suffered by DNA, also leaves a deoxyribose sugar with a missing base. Depurinations are directly repaired beginning with AP endonuclease, following the bottom half of the pathway in Figure 5–48A.

The second major repair pathway is called **nucleotide excision repair**. This mechanism can repair the damage caused by almost any large change in the structure of the DNA double helix. Such "bulky lesions" include those created by the covalent reaction of DNA bases with large hydrocarbons (such as the carcinogen benzopyrene), as well as the various pyrimidine dimers (T-T, T-C, and C-C) caused by sunlight. In this pathway, a large multienzyme complex scans the DNA for a distortion in the double helix, rather than for a specific base change. Once it finds a bulky lesion, it cleaves the phosphodiester backbone of the abnormal strand on both sides of the distortion, and a DNA helicase peels away a single-strand oligonucleotide containing the lesion. The large gap produced in the DNA helix is then repaired by DNA polymerase and DNA ligase (Figure 5–48B).

An alternative to base and nucleotide excision-repair processes is direct chemical reversal of DNA damage, and this strategy is employed for the rapid removal of certain highly mutagenic or cytotoxic lesions. For example, the alkylation lesion O^6-methylguanine has its methyl group removed by direct transfer to a cysteine residue in the repair protein itself, which is destroyed in the reaction. In another example, methyl groups in the alkylation lesions 1-methyladenine and 3-methylcytosine are "burnt off" by an iron-dependent demethylase, with release of formaldehyde from the methylated DNA and regeneration of the native base.

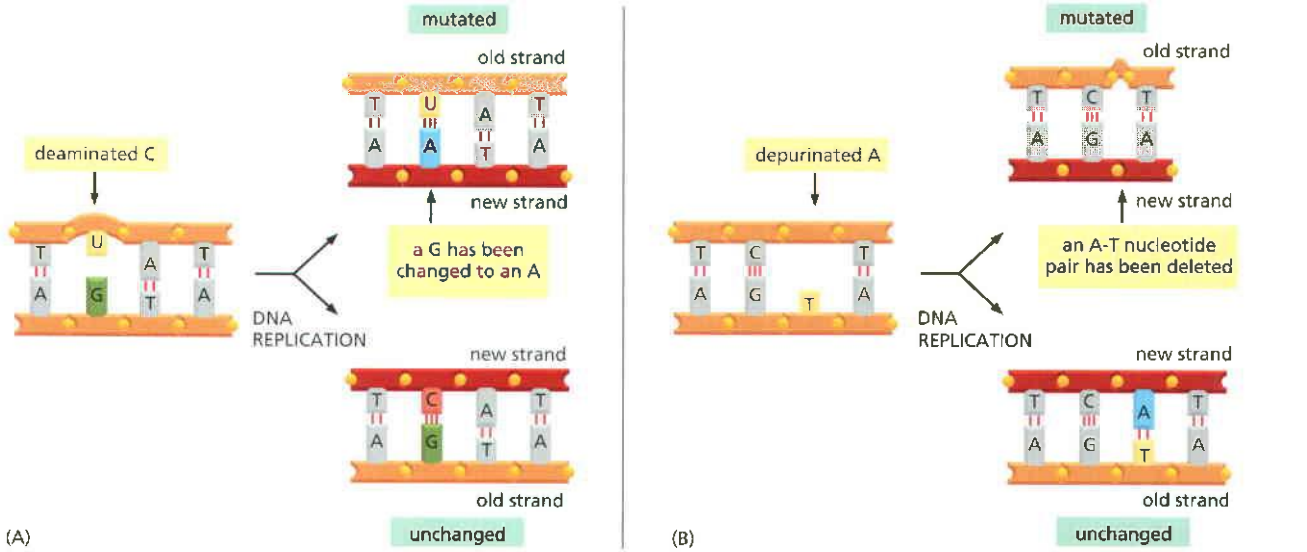


Figure 5–47 How chemical modifications of nucleotides produce mutations. (A) Deamination of cytosine, if uncorrected, results in the substitution of one base for another when the DNA is replicated. As shown in Figure 5–45, deamination of cytosine produces uracil. Uracil differs from cytosine in its base-pairing properties and preferentially base pairs with adenine. The DNA replication machinery therefore adds an adenine when it encounters a uracil on the template strand. (B) Depurination can lead to the loss of a nucleotide pair. When the replication machinery encounters a missing purine on the template strand, it may skip to the next complete nucleotide as illustrated here, thus producing a nucleotide deletion in the newly synthesized strand. Many other types of DNA damage (see Figure 5–44), if left uncorrected, also produce mutations when the DNA is replicated.

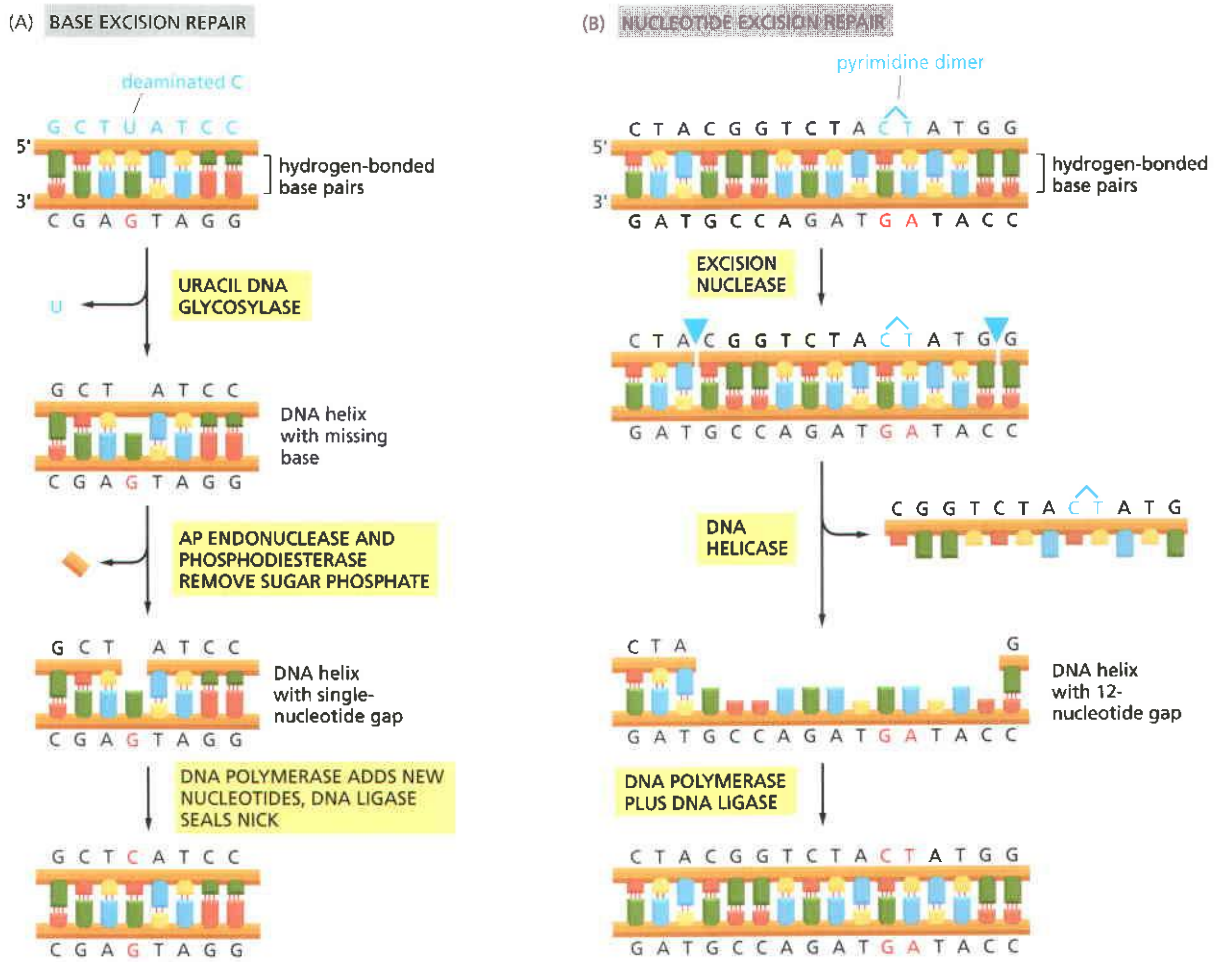


Figure 5–48 A comparison of two major DNA repair pathways. (A) *Base excision repair.* This pathway starts with a DNA glycosylase. Here the enzyme uracil DNA glycosylase removes an accidentally deaminated cytosine in DNA. After the action of this glycosylase (or another DNA glycosylase that recognizes a different kind of damage), the sugar phosphate with the missing base is cut out by the sequential action of AP endonuclease and a phosphodiesterase. (These same enzymes begin the repair of depurinated sites directly.) The gap of a single nucleotide is then filled by DNA polymerase and DNA ligase. The net result is that the U that was created by accidental deamination is restored to a C. AP endonuclease is so-named because it recognizes any site in the DNA helix that contains a deoxyribose sugar with a missing base; such sites can arise either by the loss of a purine (*ap*urinic sites) or by the loss of a pyrimidine (*ap*yrimidinic sites). (B) *Nucleotide excision repair.* In bacteria, after a multienzyme complex has recognized a lesion such as a pyrimidine dimer (see Figure 5–46), one cut is made on each side of the lesion, and an associated DNA helicase then removes the entire portion of the damaged strand. The excision repair machinery in bacteria leaves the gap of 12 nucleotides shown. In humans, once the damaged DNA is recognized, a helicase is recruited to unwind the DNA duplex locally. Next, the excision nuclease enters and cleaves on either side of the damage, leaving a gap of about 30 nucleotides. The nucleotide excision repair machinery in both bacteria and humans can recognize and repair many different types of DNA damage.

Coupling DNA Repair to Transcription Ensures That the Cell's Most Important DNA Is Efficiently Repaired

All of a cell's DNA is under constant surveillance for damage, and the repair mechanisms we have described act on all parts of the genome. However, cells have a way of directing DNA repair to the DNA sequences that are most urgently needed. They do this by linking RNA polymerase, the enzyme that transcribes DNA into RNA as the first step in gene expression, to the repair of DNA damage. RNA polymerase stalls at DNA lesions and, through the use of coupling proteins, directs the repair machinery to these sites. In bacteria, where genes are relatively short, the stalled RNA polymerase can be dissociated from the DNA, the DNA is repaired, and the gene is transcribed again from the beginning. In eucaryotes, where genes can be enormously long, a more complex reaction is used to "back up" the RNA polymerase, repair the damage, and then restart the polymerase.

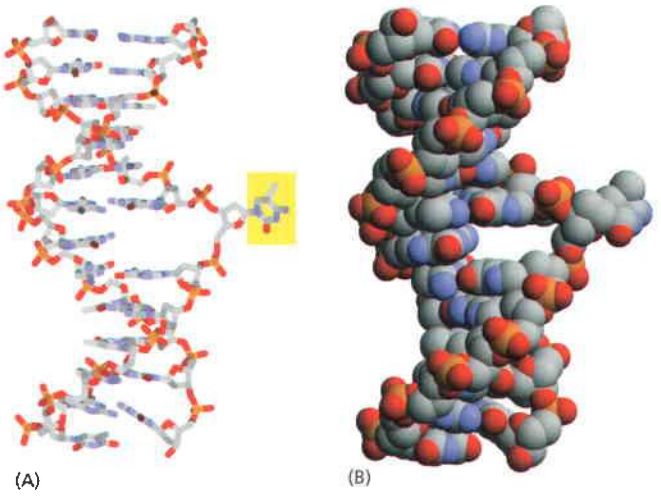


Figure 5–49 The recognition of an unusual nucleotide in DNA by base-flipping. The DNA glycosylase family of enzymes recognizes specific bases in the conformation shown. Each of these enzymes cleaves the glycosyl bond that connects a particular recognized base *(yellow)* to the backbone sugar, removing it from the DNA. (A) Stick model; (B) space-filling model.

Transcription-coupled repair works with base excision, nucleotide excision, and other repair machinery to direct repair immediately to the cell's most important DNA sequences, namely those being expressed when the damage occurs. Remarkably, this type of repair is specific for the template strand of transcribed DNA; the other strand is repaired with the same speed and efficiency as DNA that is not being transcribed at all. Transcription-coupled repair is particularly advantageous in humans, because only a small fraction of our genome is transcribed at any given time. Its importance is seen in individuals with Cockayne's syndrome, which is caused by a defect in transcription-coupled repair. These individuals suffer from growth retardation, skeletal abnormalities, progressive neural retardation, and severe sensitivity to sunlight. Most of these problems are thought to arise from RNA polymerase molecules that become permanently stalled at sites of DNA damage that lie in important genes.

The Chemistry of the DNA Bases Facilitates Damage Detection

The DNA double helix seems to be optimally constructed for repair. As noted above, it contains a backup copy of all genetic information. Equally importantly, the nature of the four bases in DNA makes the distinction between undamaged and damaged bases very clear. For example, every possible deamination event in DNA yields an "unnatural" base, which can be directly recognized and removed by a specific DNA glycosylase. Hypoxanthine, for example, is the simplest purine base capable of pairing specifically with C, but hypoxanthine is the direct deamination product of A (**Figure 5–50A**). The addition of a second amino group to hypoxanthine produces G, which cannot be formed from A by spontaneous deamination, and whose deamination product (xanthine) is likewise unique.

As discussed in Chapter 6, RNA is thought, on an evolutionary time-scale, to have served as the genetic material before DNA, and it seems likely that the genetic code was initially carried in the four nucleotides A, C, G, and U. This raises the question of why the U in RNA was replaced in DNA by T (which is 5-methyl U). We have seen that the spontaneous deamination of C converts it to U, but that this event is rendered relatively harmless by uracil DNA glycosylase. However, if DNA contained U as a natural base, the repair system would find it difficult to distinguish a deaminated C from a naturally occurring U.

A special situation occurs in vertebrate DNA, in which selected C nucleotides are methylated at specific C-G sequences that are associated with inactive genes (discussed in Chapter 7). The accidental deamination of these methylated C nucleotides produces the natural nucleotide T (Figure 5–50B) in a mismatched base pair with a G on the opposite DNA strand. To help in repairing deaminated methylated C nucleotides, a special DNA glycosylase recognizes a

mismatched base pair involving T in the sequence T-G and removes the T. This DNA repair mechanism must be relatively ineffective, however, because methylated C nucleotides are common sites for mutations in vertebrate DNA. It is striking that, even though only about 3% of the C nucleotides in human DNA are methylated, mutations in these methylated nucleotides account for about one-third of the single-base mutations that have been observed in inherited human diseases.

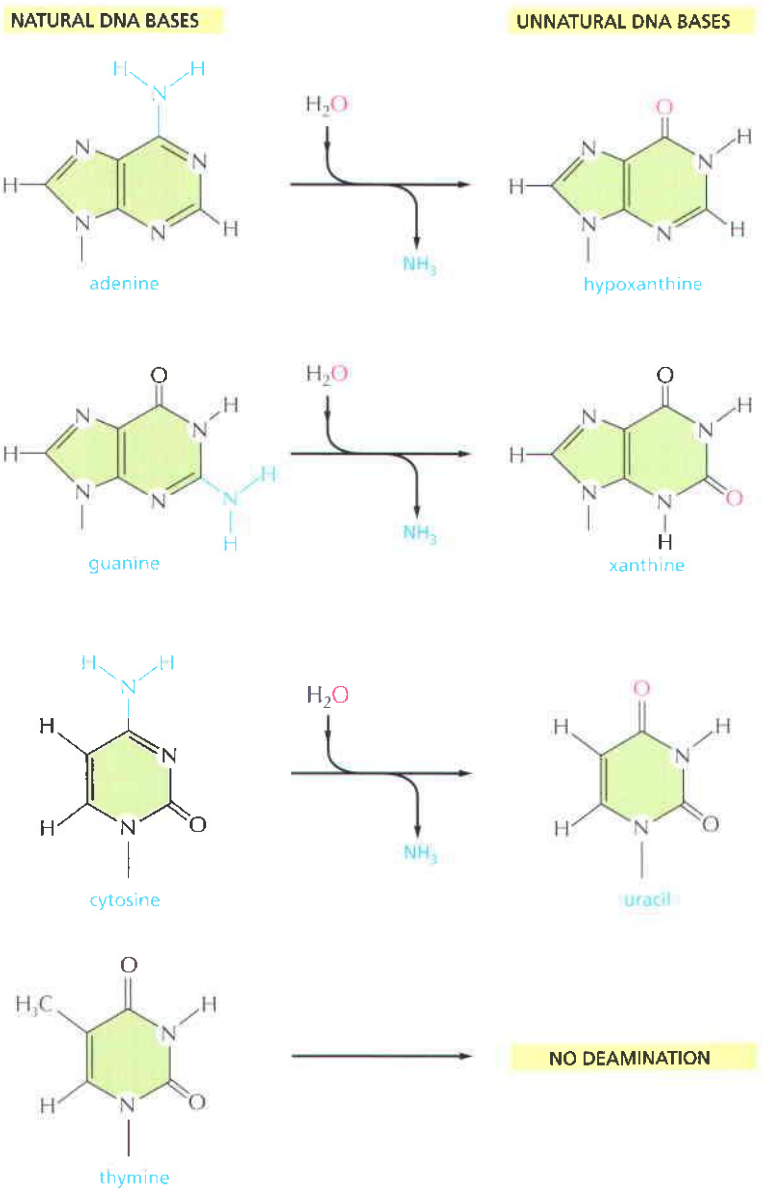


Figure 5–50 The deamination of DNA nucleotides. In each case, the oxygen atom that is added in this reaction with water is colored *red*. (A) The spontaneous deamination products of A and G are recognizable as unnatural when they occur in DNA and thus are readily recognized and repaired. The deamination of C to U was also illustrated in Figure 5–45; T has no amino group to remove. (B) About 3% of the C nucleotides in vertebrate DNAs are methylated to help in controlling gene expression (discussed in Chapter 7). When these 5-methyl C nucleotides are accidentally deaminated, they form the natural nucleotide T. However, this T will be paired with a G on the opposite strand, forming a mismatched base pair.

Special DNA Polymerases Are Used in Emergencies to Repair DNA

If a cell's DNA is heavily damaged, the repair mechanisms that we have discussed are often insufficient to cope with it. In these cases, a different strategy is called into play, one that entails some risk to the cell. The highly accurate replicative DNA polymerases stall when they encounter damaged DNA, and in emergencies cells employ versatile but less accurate back-up polymerases to replicate through the DNA damage.

Human cells contain more than 10 such DNA polymerases, some of which can recognize a specific type of DNA damage and specifically add the nucleotide required to restore the initial sequence. Others make only "good guesses," especially when the template base has been extensively damaged. These enzymes are not as accurate as the normal replicative polymerases when they copy a normal DNA sequence. For one thing, the backup polymerases lack exonucleolytic proofreading activity; in addition, many are much less discriminating than the replicative polymerase in choosing which nucleotide to initially incorporate. Presumably, for this reason, each such polymerase molecule is given a chance to add only one or a few nucleotides. Although the details of these fascinating reactions are still being worked out, they provide elegant testimony to the care with which organisms maintain the integrity of their DNA.

Double-Strand Breaks Are Efficiently Repaired

An especially dangerous type of DNA damage occurs when both strands of the double helix are broken, leaving no intact template strand to enable accurate repair. Ionizing radiation, replication errors, oxidizing agents, and other metabolites produced in the cell cause breaks of this type. If these lesions were left unrepaired, they would quickly lead to the breakdown of chromosomes into smaller fragments and to the loss of genes when the cell divides. However, two distinct mechanisms have evolved to ameliorate this type of damage (**Figure 5–51**). The simplest to understand is **nonhomologous end-joining**, in which the broken ends are simply brought together and rejoined by DNA ligation, generally with the loss of one or more nucleotides at the site of joining (**Figure 5–52**). This end-joining mechanism, which might be seen as a "quick and dirty" solution to the repair of double-strand breaks, is common in mammalian somatic cells. Although a change in the DNA sequence (a mutation) results at the site of breakage, so little of the mammalian genome codes for proteins that this mechanism is apparently an acceptable solution to the problem of rejoining broken

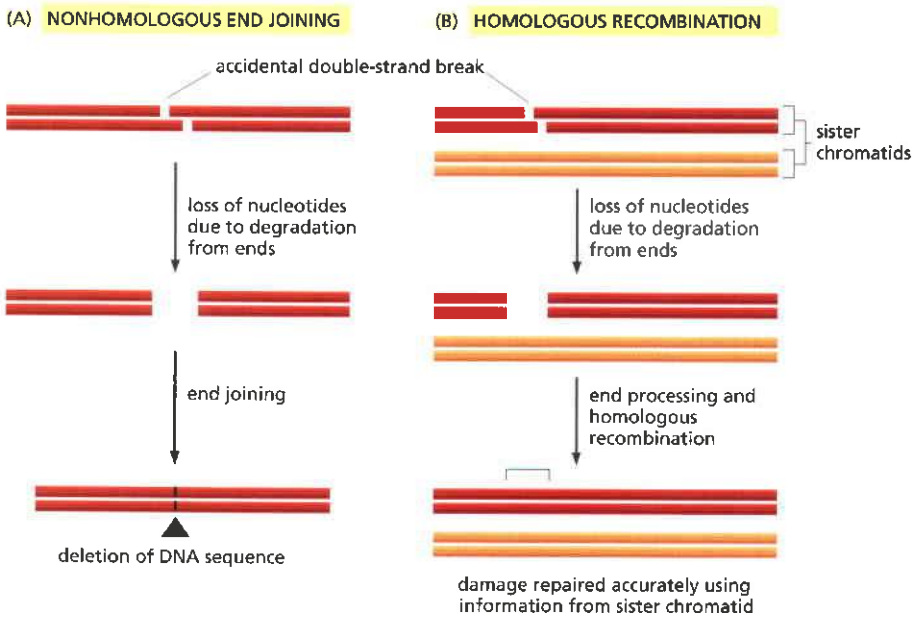


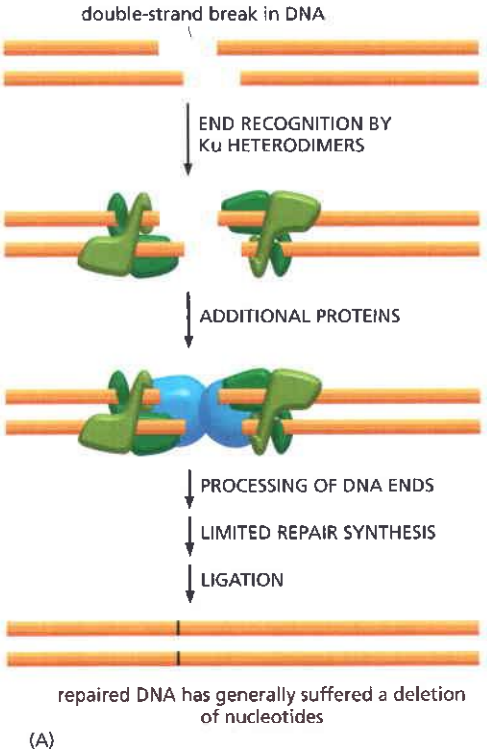
Figure 5–51 Two different ways to repair double-strand breaks. (A) Nonhomologous end-joining alters the original DNA sequence when repairing a broken chromosome. These alterations can be either deletions (as shown) or short insertions. (B) Repairing double-strand breaks by homologous recombination is more difficult to accomplish, but this type of repair restores the original DNA sequence. It typically takes place after the DNA has been duplicated but before the cell has divided. Details of the homologous recombination pathway will be presented later (see Figure 5–61).

chromosomes. By the time a human reaches the age of 70, the typical somatic cell contains over 2000 such "scars", distributed throughout its genome, representing places where DNA has been inaccurately repaired by nonhomologous end-joining. As previously discussed, the specialized structure of telomeres prevents the natural ends of chromosomes from being mistaken for broken DNA and repaired.

A much more accurate type of double-strand break repair occurs in newly replicated DNA (Figure 5–51B). Here, the DNA is repaired using the sister chromatid as a template. This reaction is an example of *homologous recombination*, and we consider its mechanism later in this chapter. Most organisms employ both nonhomologous end-joining and homologous recombination to repair double-strand breaks in DNA. Nonhomologous end-joining predominates in humans; homologous recombination is used only during and shortly after DNA replication (in S and G_2 phases), when sister chromatids are available to serve as templates.

DNA Damage Delays Progression of the Cell Cycle

We have just seen that cells contain multiple enzyme systems that can recognize and repair many types of DNA damage. Because of the importance of maintaining intact, undamaged DNA from generation to generation, eucaryotic cells have an additional mechanism that maximizes the effectiveness of their DNA repair enzymes: they delay progression of the cell cycle until DNA repair is complete. As discussed in detail in Chapter 17, the orderly progression of the cell cycle is maintained through the use of *checkpoints* that ensure the completion of one step before the next step can begin. At several of these cell-cycle checkpoints, the cycle stops if damaged DNA is detected. Thus, in mammalian cells, the presence of DNA damage can block entry from G_1 into S phase, it can slow S phase once it has begun, and it can block the transition from S phase to M phase. These delays facilitate DNA repair by providing the time needed for the repair to reach completion.



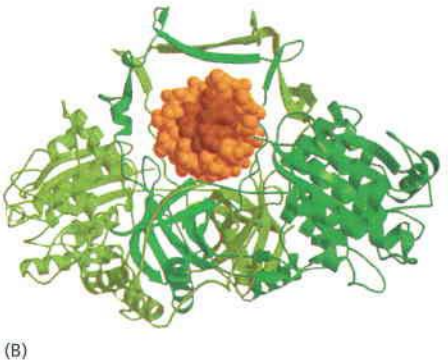


Figure 5–52 Nonhomologous end-joining. (A) A central role is played by the Ku protein, a heterodimer that grasps the broken chromosome ends. The additional proteins shown are needed to hold the broken ends together while they are processed and eventually joined covalently. (B) Three-dimensional structure of a Ku heterodimer bound to the end of a duplex DNA fragment. The Ku protein is also essential for V(D)J joining, a specific recombination process through which antibody and T cell receptor diversity is generated in developing B and T cells (discussed in Chapter 25). V(D)J joining and nonhomologous end-joining show many similarities in mechanism but the former relies on specific double-strand breaks produced deliberately by the cell. (B, from J.R. Walker, R.A. Corpina and J. Goldberg, *Nature* 412:607–614, 2001. With permission from Macmillan Publishers Ltd.)

DNA damage also results in an increased synthesis of some DNA repair enzymes. The importance of special signaling mechanisms that respond to DNA damage is indicated by the phenotype of humans who are born with defects in the gene that encodes the *ATM protein*. These individuals have the disease *ataxia telangiectasia (AT)*, the symptoms of which include neurodegeneration, a predisposition to cancer, and genome instability. The ATM protein is a large kinase needed to generate the intracellular signals that produce a response to many types of spontaneous DNA damage, and individuals with defects in this protein therefore suffer from the effects of unrepaired DNA lesions.

Summary

Genetic information can be stored stably in DNA sequences only because a large set of DNA repair enzymes continuously scan the DNA and replace any damaged nucleotides. Most types of DNA repair depend on the presence of a separate copy of the genetic information in each of the two strands of the DNA double helix. An accidental lesion on one strand can therefore be cut out by a repair enzyme and a corrected strand resynthesized by reference to the information in the undamaged strand.

Most of the damage to DNA bases is excised by one of two major DNA repair pathways. In base excision repair, the altered base is removed by a DNA glycosylase enzyme, followed by excision of the resulting sugar phosphate. In nucleotide excision repair, a small section of the DNA strand surrounding the damage is removed from the DNA double helix as an oligonucleotide. In both cases, the gap left in the DNA helix is filled in by the sequential action of DNA polymerase and DNA ligase, using the undamaged DNA strand as the template. Some types of DNA damage can be repaired by a different strategy—the direct chemical reversal of the damage—which is carried out by specialized repair proteins.

Other critical repair systems—based on either nonhomologous end-joining or homologous recombination—reseal the accidental double-strand breaks that occur in the DNA helix. In most cells, an elevated level of DNA damage causes a delay in the cell cycle via checkpoint mechanisms, which ensure that DNA damage is repaired before a cell divides.

HOMOLOGOUS RECOMBINATION

In the two preceding sections, we discussed the mechanisms that allow the DNA sequences in cells to be maintained from generation to generation with very little change. In this section, we further explore one of these mechanisms, *homologous recombination*. Although crucial for accurately repairing double-strand breaks (see Figure 5–51B) and other types of DNA damage, homologous recombination, as we shall see, can also rearrange DNA sequences. These rearrangements often alter the particular versions of genes present in an individual genome, as well as the timing and the level of their expression. In a population, the type of genetic variation produced by this and other types of genetic recombination is crucial for facilitating the evolution of organisms in response to a changing environment.

Homologous Recombination Has Many Uses in the Cell

In **homologous recombination** (also known as *general recombination*), genetic exchange takes place between a pair of homologous DNA sequences, that is, DNA sequences similar or identical in nucleotide sequence. Homologous recombination has many uses in the cell, but three are of paramount importance. The most widespread use is in accurately repairing double-strand breaks, as introduced in the previous section (Figure 5–51B). Although double-strand breaks can result from radiation and reactive chemicals, many arise from DNA replication forks that become stalled or broken. This application of homologous

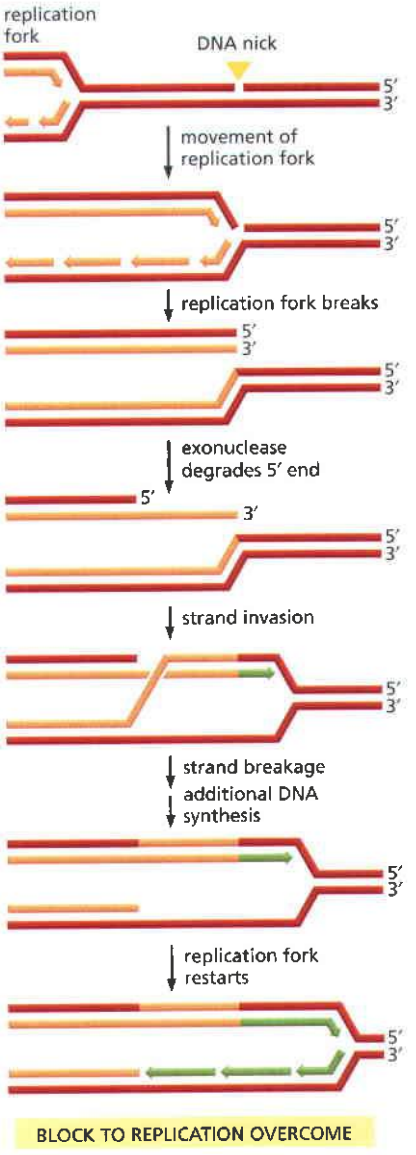


Figure 5–53 Repair of a broken replication fork by homologous recombination. When a moving replication fork encounters a single-strand break, it will collapse but can be repaired by homologous recombination. As shown, the initial strand invasion requires a free 3′ end generated by a nuclease that degrades the 5′ end of the complementary strand. Recombination then begins with strand invasion, as described in detail in subsequent figures. Arrowheads represent 3′ ends of DNA strands. Green strands represent the new DNA synthesis that takes place after the replication fork has broken. Note that, in this mechanism, the fork moves past the site that was nicked on the original template by using an undamaged copy of the site as its template. (Adapted from M.M. Cox, *Proc. Natl Acad. Sci. U.S.A.* 98:8173–8180, 2001. With permission from National Academy of Sciences.)

recombination is essential for every proliferating cell, because accidents occur during nearly every round of DNA replication.

Many types of events can cause the replication fork to break during the replication process. Consider just one example: a single-strand nick or gap in the parental DNA helix just ahead of a replication fork. When the fork reaches this lesion, it falls apart—resulting in one broken and one intact daughter chromosome. However, a series of recombination reactions, which can begin with a *strand invasion* process that triggers DNA synthesis by DNA polymerase, can flawlessly repair the broken chromosome (**Figure 5–53**).

In addition, homologous recombination is used to exchange bits of genetic information between two different chromosomes to create new combinations of DNA sequences in each chromosome. The potential evolutionary benefit of this type of gene mixing is that it creates an array of new, perhaps beneficial, combinations of genes. During meiosis in fungi, plants, and animals, homologous recombination also plays an important mechanical role in assuring accurate chromosome segregation. In this section, we consider only the universal roles of homologous recombination, those of repairing DNA damage and of mediating genetic exchange. Its more specialized mechanical role in chromosome segregation during meiosis will be discussed in Chapter 21.

Homologous Recombination Has Common Features in All Cells

The current view of homologous recombination as a critical DNA repair mechanism in all cells evolved slowly from its initial discovery as a key component in the specialized process of meiosis in plants and animals. The subsequent recognition that homologous recombination also occurs in less complex unicellular organisms made it much more amenable to molecular analyses. Thus, most of what we know about the biochemistry of genetic recombination was originally derived from studies of bacteria, especially of *E. coli* and its viruses, as well as from experiments with simple eucaryotes such as yeasts. For these organisms with short generation times and relatively small genomes, it was possible to isolate a large set of mutants with defects in their recombination processes. The protein altered in each mutant was then identified and, ultimately, studied biochemically. More recently, close relatives of these proteins have been discovered and extensively characterized in *Drosophila*, mice, and humans. These studies reveal that the fundamental processes that catalyze homologous recombination are common to all cells, as we shall now discuss.

DNA Base-Pairing Guides Homologous Recombination

The hallmark of homologous recombination is that it takes place only between DNA duplexes that have extensive regions of sequence similarity (homology). Not surprisingly, base-pairing underlies this requirement, and two DNA duplexes that are undergoing homologous recombination "sample" each other's DNA sequence by engaging in extensive base-pairing between a single strand

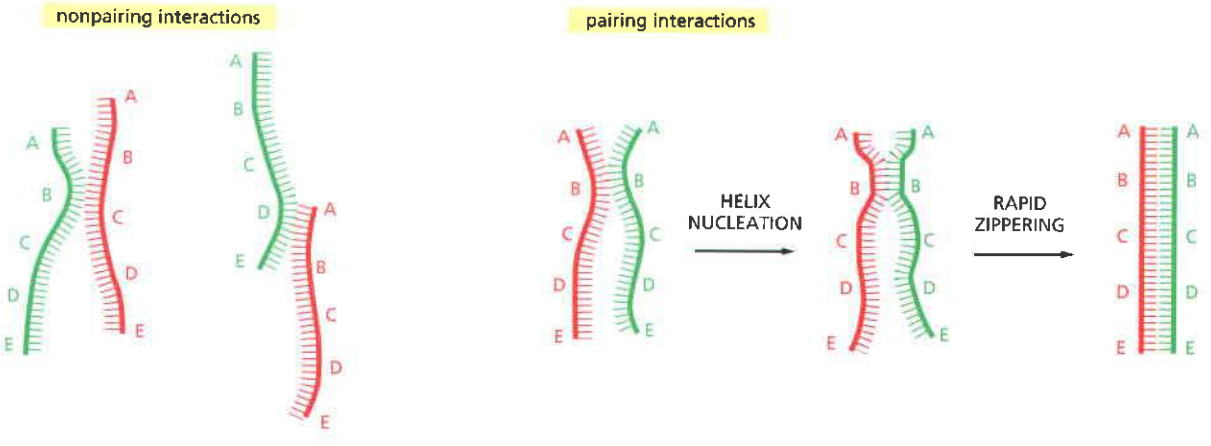
from one DNA duplex and the complementary single strand from the other. The match need not be perfect, but it must be very close for homologous recombination to succeed.

In its simplest form, this type of base-pairing interaction can be mimicked in a test tube by allowing a DNA double helix to re-form from its separated single strands. This process, called *DNA renaturation* or **hybridization**, occurs when a rare random collision juxtaposes complementary nucleotide sequences on two matching DNA single strands, allowing the formation of a short stretch of double helix between them. This relatively slow helix nucleation step is followed by a very rapid "zippering" step, as the region of double helix is extended to maximize the number of base-pairing interactions (**Figure 5–54**).

Formation of a new double helix in this way requires that the annealing strands be in an open, unfolded conformation. For this reason, *in vitro* hybridization reactions are performed either at high temperature or in the presence of an organic solvent such as formamide; these conditions "melt out" the short hairpin helices that result from the base-pairing interactions that occur within a single strand that folds back on itself. Most cells cannot survive such harsh conditions and instead use single-strand DNA-binding proteins (see p. 273) to melt out the hairpin helices. Single-strand DNA-binding proteins are essential for DNA replication (as described earlier) as well as for homologous recombination; they bind tightly and cooperatively to the sugar-phosphate backbone of all single-stranded DNA regions of DNA, holding them in an extended conformation with the bases exposed (see Figures 5–16 and 5–17). In this extended conformation, a DNA single strand can base-pair efficiently either with a nucleoside triphosphate molecule (in DNA replication) or with a complementary section of another DNA single strand (as part of a genetic recombination process).

DNA hybridization creates a region of DNA helix formed from strands that originate from two different DNA molecules. The formation of such a region, known as a *heteroduplex*, is an essential step in any homologous recombination process. Because the vast majority of DNA inside the cell is double-stranded, the "test tube" model of DNA hybridization cannot fully explain how this process occurs in the cell. Indeed, special mechanisms are required to start homologous recombination between two double-stranded DNA molecules of similar nucleotide sequence. Crucial to these mechanisms are proteins that allow DNA hybridization to occur in the cell through *strand invasion*—the pairing of a region of single-stranded DNA with a complementary strand in a different DNA double helix—as we describe next.

Figure 5–54 DNA hybridization. DNA double helices re-form from their separated strands in a reaction that depends on the random collision of two complementary DNA strands. The vast majority of such collisions are not productive, as shown on the left, but a few result in a short region where complementary base pairs have formed (helix nucleation). A rapid zippering then leads to the formation of a complete double helix. Through this trial-and-error process, a DNA strand will find its complementary partner even in the midst of millions of nonmatching DNA strands.



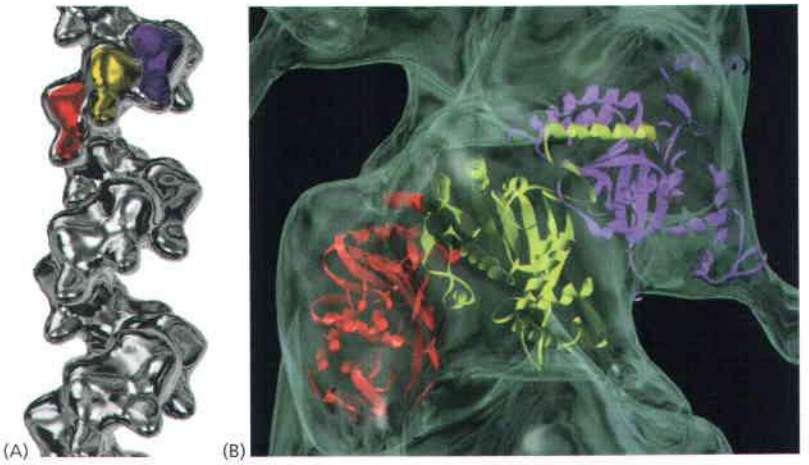


Figure 5–55 The structure of the RecA and Rad51 protein–DNA filaments. (A) The Rad51 protein, a human homolog of the bacterial RecA protein, is bound to a DNA single strand. Three successive protein monomers in this helical filament are colored. (B) A short section of the RecA filament, with the three-dimensional structure of the protein fitted to the image of the filament determined by electron microscopy. There are about six RecA monomers per turn of the helix, holding 18 nucleotides of single-stranded DNA that is stretched out by the protein. The exact path of the DNA in this structure is not known. (A, courtesy of Edward Egelman; B, from X. Yu et al., *J. Mol. Biol.* 283:985–992, 1998. With permission from Academic Press.)

The RecA Protein and its Homologs Enable a DNA Single Strand to Pair with a Homologous Region of DNA Double Helix

Because extensive base-pair interactions cannot occur between two intact DNA double helices, the DNA hybridization that is critical for homologous recombination can begin only after a DNA strand from one DNA helix is freed from pairing with its complementary strand, thereby making its nucleotide available for pairing with a second DNA helix. In the example that was previously illustrated in Figure 5–53, this free single strand is formed when a replication fork encounters a DNA nick, falls apart (creating a new double-stranded end), and an exonuclease degrades the 5′ end at the break, leaving an unpaired single strand at its 3′ end. In other applications of homologous recombination, single stranded regions are generated in similar ways, as we discuss later.

The single-strand at the 3′ DNA end is acted upon by several specialized proteins that direct it to invade a homologous DNA duplex. Of central importance is the **RecA protein**, its name in *E.coli*, and its homolog **Rad51**, its name in virtually all eucaryotic organisms (**Figure 5–55**). Like a single-strand DNA-binding protein, the RecA type of protein binds tightly and in long cooperative clusters to single-stranded DNA forming a nucleoprotein filament. Because each RecA monomer has more than one DNA-binding site, a RecA filament can hold a single strand and a double helix together (**Figure 5–56**). This arrangement allows RecA to catalyze a multistep *DNA synapsis* reaction that occurs between a DNA double helix and a homologous region of single-stranded DNA. In the first step, the RecA protein intertwines the DNA single strand and the DNA duplex in a sequence-independent manner. Next, the DNA single strand "searches" the duplex for homologous sequences. Exactly how this searching and eventual recognition occurs is not understood, but it may involve transient base pairs formed between the single strand and bases that flip out from the duplex DNA (see Figure 5–49). Once a homologous sequence has been located, a strand invasion occurs: the single strand displaces one strand of the duplex as

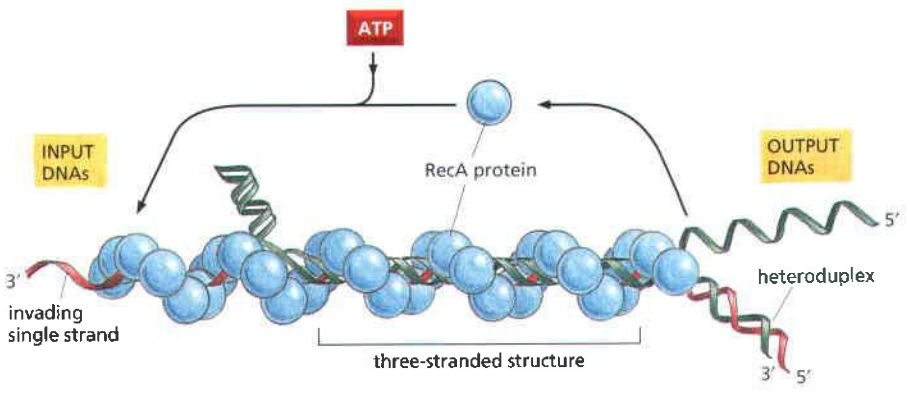


Figure 5–56 DNA synapsis catalyzed by the RecA protein. *In vitro* experiments show that several types of complex are formed between a DNA single strand *(red)* covered with RecA protein *(blue)* and a DNA double helix *(green)*. First a non-base-paired complex is formed, which is converted into a "joint molecule" as a homologous sequence is found. This complex is dynamic and spins out a DNA heteroduplex (one strand *green* and the other strand *red*) plus a displaced single strand from the original helix *(green)*. Thus, the structure shown in this diagram migrates to the left, reeling in the "input DNAs" while producing the "output DNAs." (Adapted from S.C. West, *Annu. Rev. Biochem.* 61:603–640, 1992. With permission from Annual Reviews.)

it forms conventional base pairs with the other strand. The result is a **heteroduplex**—a region of DNA double helix formed by the pairing of two DNA strands that were initially part of two different DNA molecules (see Figure 5–56).

The search for homology and the invasion of a single strand into a DNA duplex are the critical reactions that initiate homologous recombination. They require, in addition to RecA-like proteins and single-strand binding proteins, several proteins with specialized functions. For example, Rad52 protein displaces the single-strand binding proteins allowing the binding of Rad51 molecules, and in addition, promotes the annealing of complementary single-strands (Figure 5–57).

The short heteroduplex region formed, where the invading single strand has paired with its complementary strand in the DNA duplex, is often greatly enlarged by a process called branch migration, as we now discuss.

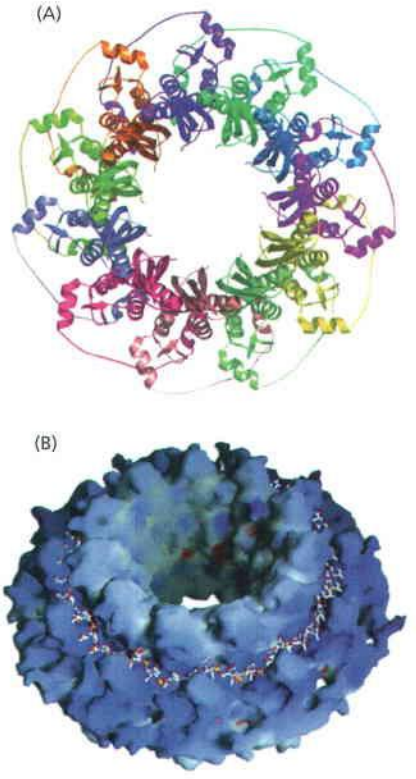


Figure 5–57 Structure of a portion of the Rad52 protein. (A) This doughnut-shaped structure is composed of 11 subunits. (B) Single-stranded DNA has been modeled into the deep groove running along the protein surface. The bases of the DNA are exposed, a configuration that is proposed to mediate the annealing of two complementary single strands. (From M.R. Singleton et al., *Proc. Natl Acad. Sci. U.S.A.* 99:13492–13497, 2002. With permission from the National Academy of Sciences.)

Branch Migration Can either Enlarge Heteroduplex Regions or Release Newly Synthesized DNA as a Single Strand

Once a strand invasion reaction has occurred, the point of strand exchange (the "branch point") can move through a process called *branch migration* (**Figure 5–58**). In this reaction, an unpaired region of one of the single strands displaces a paired region of the other single strand, moving the branch point without changing the total number of DNA base pairs. Although spontaneous branch migration can occur, it proceeds equally in both directions, so it makes little net progress (Figure 5–58A). Specialized DNA helicases, however, can catalyze unidirectional branch migration, readily producing a region of heteroduplex DNA that can be thousands of base pairs long (Figure 5–58B).

In a related reaction, DNA synthesis catalyzed by DNA polymerase can drive a unidirectional branch migration process through which the newly synthesized DNA is displaced as a single strand, mimicking the way that a newly synthesized RNA chain is released by RNA polymerase. This form of DNA synthesis appears to be used in several homologous recombination processes, including the double-strand break repair processes to be described next.

Homologous Recombination Can Flawlessly Repair Double-Stranded Breaks in DNA

Earlier in this chapter, we discussed the dire problems created by double-strand breaks in DNA, and we saw that cells can repair these breaks in two ways. Nonhomologous end-joining (see Figure 5–51) occurs without a template and creates

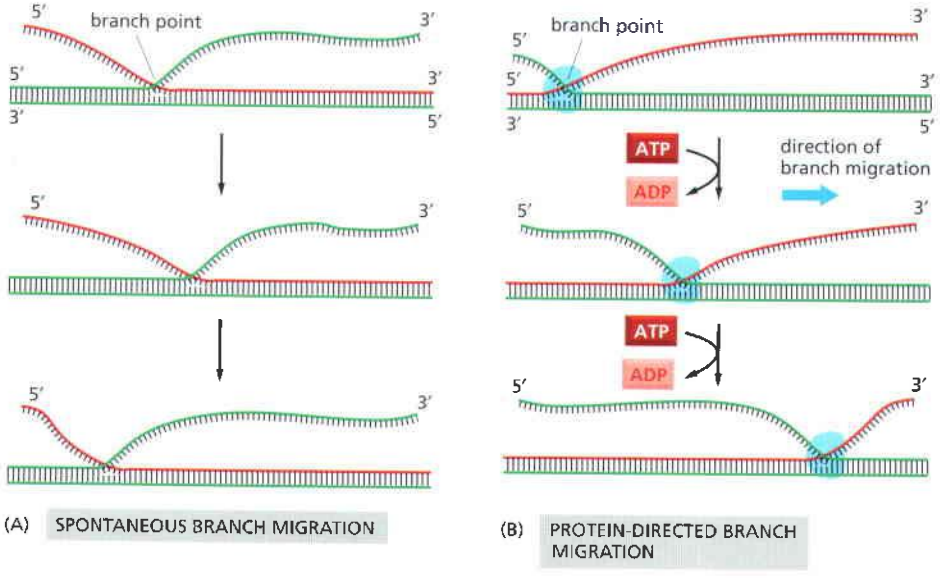


Figure 5–58 Two types of DNA branch migration observed in experiments *in vitro*. (A) Spontaneous branch migration is a back-and-forth, random-walk process, and it therefore makes little progress over long distances. (B) Protein-directed branch migration requires energy and it moves the branch point at a uniform rate in one direction.

a mutation at the site at which two DNA duplexes are joined. It can also inadvertently join together segments from two different chromosomes creating chromosome translocations, many of which have serious consequences for the cell. In contrast to nonhomologous end-joining, homologous recombination can repair double-stranded breaks accurately, without any loss or alteration of nucleotides at the site of repair (**Figure 5–59**). In most cells, recombination-mediated double-strand break repair occurs only after the cell has replicated its DNA, when one nearby daughter DNA duplex can serve as the template for repair of the other.

Homologous recombination can also be used to repair many other types of DNA damage, making it perhaps the most versatile DNA repair mechanism available to the cell; the "all-purpose" nature of recombinational repair probably explains why its mechanism and the proteins that carry it out have been conserved in virtually all cells on Earth.

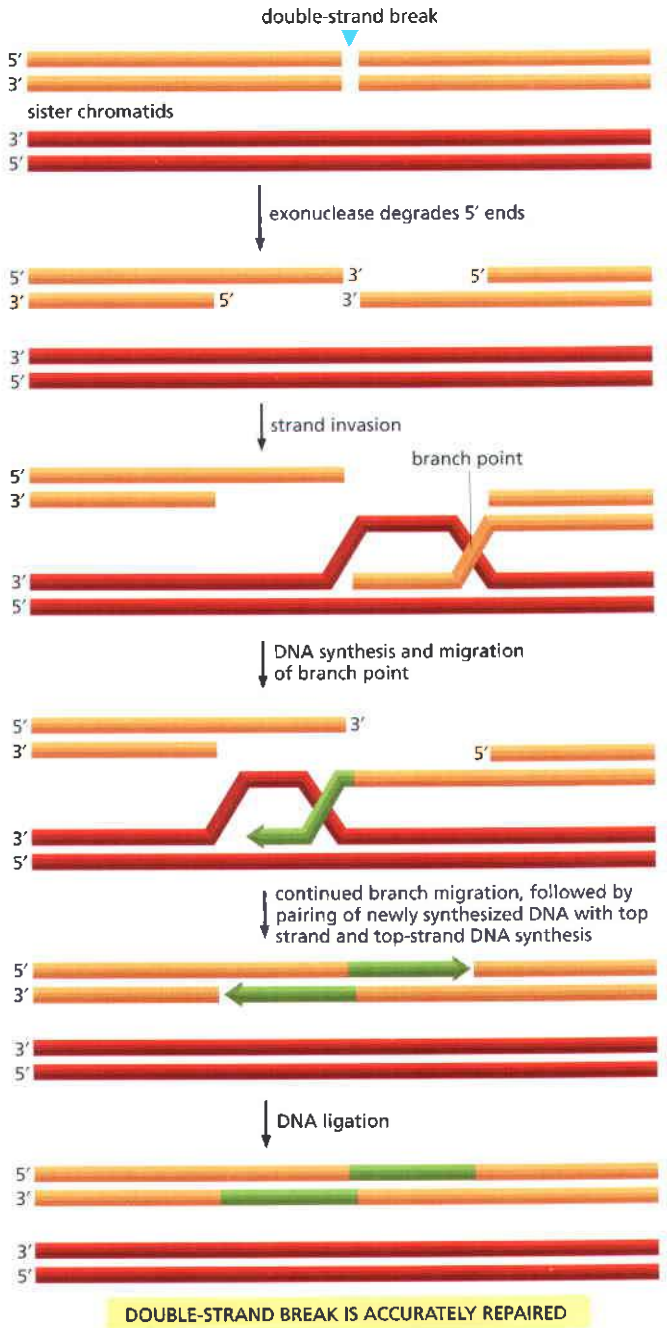


Figure 5–59 Mechanism of double-strand break repair by homologous recombination. This is the preferred method for repairing DNA double-strand breaks that arise shortly after the DNA has been replicated and the two sister chromatids are still held together. This reaction uses many of the same proteins as that of Figure 5–53 and proceeds through the same basic intermediate steps. In general, homologous recombination can be regarded as a flexible series of reactions, with the exact pathway differing from one case to the next. For example, the length of the repair "patch" can vary considerably depending on the extent of 5′ processing, branch migration, and new DNA synthesis. (See M. McVey, J. LaRocque, M.D. Adams, and J. Sekelsky, *Proc. Natl Acad. Sci. U.S.A.* 101:15694–15699, 2004.)

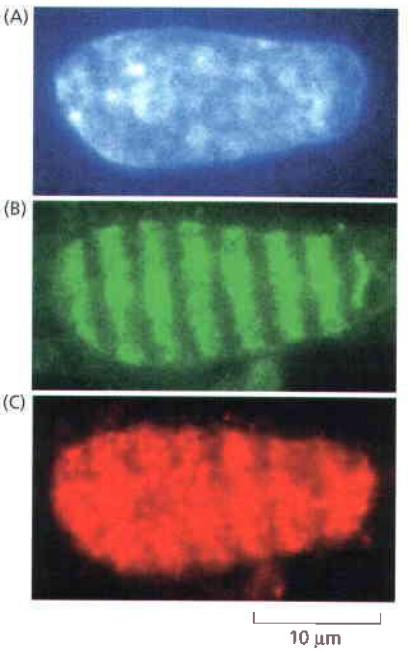


Figure 5–60 Experiment demonstrating the rapid localization of repair proteins to DNA double-strand breaks. Human fibroblasts were x-irradiated to produce DNA double-strand breaks. Before the x-rays struck the cells, they were passed through a microscopic grid with x-ray-absorbing "bars" spaced 1 μm apart. This produced a striped pattern of DNA damage, allowing a comparison of damaged and undamaged DNA in the same nucleus. (A) Total DNA in a fibroblast nucleus stained with the dye DAPI. (B) Sites of new DNA synthesis indicated by incorporation of BudR (a thymine analog) and subsequent staining with FITC-coupled antibodies to BudR. (C) Localization of the Mre11 complex to damaged DNA as visualized by Texas-red complexed to antibodies against the Mre11 subunit. It has been proposed that the Mre11 complex initially recognizes double-strand breaks in the cell and then mobilizes additional proteins to repair the breaks through homologous recombination (see Figure 5–59). (A), (B) and (C) were processed 30 minutes after x-irradiation. (From B.E. Nelms et al., *Science* 280:590–592, 1998. With permission from AAAS.)

Cells Carefully Regulate the Use of Homologous Recombination in DNA Repair

Although homologous recombination neatly solves the problem of accurately repairing double-strand breaks and other types of DNA damage, it does present some dangers to the cell and therefore must be tightly regulated. For example, the DNA sequence in one chromosomal homolog can be rendered non-functional by "repairing" it using the other chromosomal homolog as the template. A *loss of heterozygosity* of this type is frequently a critical step in the formation of cancers (as discussed in Chapter 20), and cells have poorly understood mechanisms to minimize it. Although relatively rare in normal cells, loss of heterozygosity can be viewed as an unfortunate side effect of the versatility of homologous recombination.

Another type of control over recombination repair, which is found in nearly all eucaryotic cells, prevents recombination-based "repair" in the absence of DNA damage. The enzymes that catalyze recombination repair are normally made at relatively high levels in eucaryotes and are dispersed throughout the nucleus. In response to DNA damage, they rapidly converge on the sites of damage and eventually form "repair factories" where many DNA lesions are apparently brought together and repaired (**Figure 5–60**). This rapid mobilization of repair proteins to DNA damage is tightly controlled by the cell and requires a series of additional proteins. Two of these, the *Brca1* and *Brca2* proteins, were first discovered because mutations in their genes lead to a greatly increased frequency of breast cancer. Whereas the removal of a protein essential for homologous recombination (such as the human Rad51 protein) will kill a cell, an alteration in an accessory protein can lead to inefficient repair. The subsequent accumulation of DNA damage can, in a small proportion of cells, give rise to a cancer. Brca2 binds to the Rad51 protein preventing its polymerization on DNA and thereby maintaining it in an inactive form. It is thought that Brca2 helps to bring Rad51 protein rapidly to sites of damage and, once in place, to release it in its active form.

In Chapter 20, we shall see that both too much and too little homologous recombination can lead to cancer in humans, the former through an enhanced loss of heterozygosity and the latter through an increased mutation rate caused by inefficient DNA repair. Clearly, a delicate balance has evolved that keeps this process in check on undamaged DNA, while still allowing it to act efficiently and rapidly on DNA lesions as soon as they arise.

Bacteria also carefully regulate their recombination and other DNA repair enzymes, but this occurs largely by controlling their intracellular levels. In response to severe DNA damage, *E. coli* increases the transcription of many DNA repair enzymes as part of the so-called *SOS response*. These include nucleotide excision repair enzymes, error-prone DNA polymerases that can use damaged DNA as a template, and proteins that mediate homologous recombination.

Studies of mutant bacteria deficient in different parts of the SOS response demonstrate that the newly synthesized proteins have two effects. First, as would be expected, the induction of these additional DNA repair enzymes increases cell survival after DNA damage. Second, several of the induced proteins transiently increase the mutation rate by increasing the number of errors made in copying DNA sequences. The errors are caused by the production of low-fidelity DNA polymerases that can efficiently use damaged DNA as a template for DNA synthesis. While this "error-prone" DNA repair can be harmful to individual bacterial cells, it is presumed to be advantageous in the long term because it produces a burst of genetic variability in the bacterial population that increases the likelihood of a mutant cell arising that is better able to survive in the altered environment.

Holliday Junctions Are Often Formed During Homologous Recombination Events

Homologous recombination can be viewed as a group of related reactions that use single-strand invasion, branch migration, and limited DNA synthesis to exchange DNA between two double helices of similar nucleotide sequence. Having discussed its role in accurately repairing damaged DNA, we now turn to homologous recombination as a means to generate DNA molecules of novel sequence. During this process a special DNA intermediate often forms that contains four DNA strands shared between two DNA helices. In this key intermediate, known as a **Holliday junction**, or *cross-strand exchange*, two DNA strands switch partners between two double helices. The Holliday junction can adopt multiple conformations, and a special set of recombination proteins binds to, and thereby stabilizes, the open, symmetric isomer (**Figure 5–61**). By using the energy of ATP hydrolysis to coordinate two branch migration reactions, these proteins can move the point at which the two DNA helices are joined rapidly along the two helices (**Figure 5–62**).

The four-stranded DNA structures produced by homologous recombination are only transiently present in cells. Thus, to regenerate two separate DNA helices, and thus end the recombination process, the strands connecting the two helices in a Holliday junction must eventually be cut, a process referred to as *resolution*. In bacteria, where we understand this process the best, a specialized endonuclease (called RuvC) cleaves the Holliday junctions leaving nicks in the DNA that DNA ligase can seal easily. However, during the meiotic processes that produce germ cells in eucaryotes (sperm and egg in animals), the resolution mechanisms appear to be much more complicated.

As we discuss in Chapter 21, extensive homologous recombination occurs as an integral part of the process whereby chromosomes are parceled out to germ cells during meiosis. Both chromosome *crossing over* and *gene conversion* result from these recombination events, producing hybrid chromosomes that contain genetic information from both the maternal and paternal homologs (**Figure 5–63**A).

As we shall see next, in meiosis the crossing over and gene conversion are both generated by homologous recombination mechanisms that, at their core, resemble those used to repair double-strand breaks.

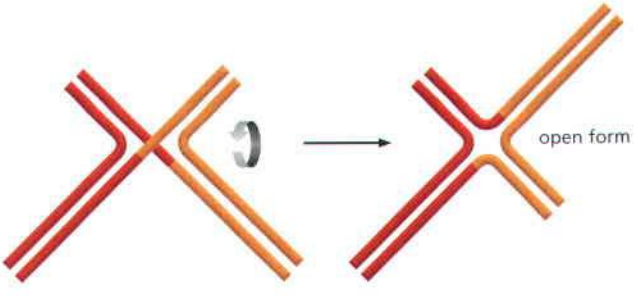


Figure 5–61 A Holliday junction. The initially formed structure is usually drawn with two crossing (inside) strands and two noncrossing (outside) strands. An isomerization of the Holliday junction produces an open, symmetrical structure <CTAG>. This is the form that is bound by the RuvA and RuvB proteins (see Figure 5–62). The Holliday junction is named for the scientist who first proposed its formation.

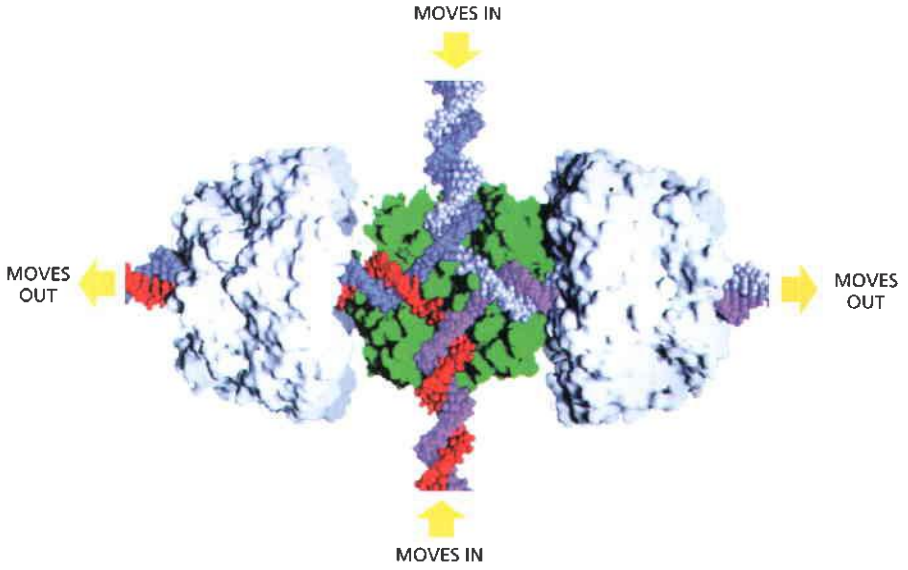


Figure 5–62 Enzyme-catalyzed double branch migration at a Holliday junction. In *E. coli*, a tetramer of the RuvA protein *(green)* and two hexamers of the RuvB protein *(pale gray)* bind to the open form of the junction. The RuvB protein, which resembles the hexameric helicases used in DNA replication (Figure 5–15), uses the energy of ATP hydrolysis to move the crossover point rapidly along the paired DNA helices, extending the heteroduplex region as shown. (Image courtesy of P. Artymiuk; modified from S.C. West, *Cell* 94:699–701, 1998. With permission from Elsevier.)

Meiotic Recombination Begins with a Programmed Double Strand Break

Homologous recombination in meiosis starts with a bold stroke: a specialized protein (called Spo11 in budding yeast) breaks both strands of the DNA double helix in one of the recombining chromosomes. Like a topoisomerase, the reaction of Spo11 with DNA leaves the protein covalently bound to the broken DNA (see Figure 5–22). A specialized nuclease then rapidly processes the ends bound by Spo11, removing the protein and leaving protruding 3′ single-strand ends. At this point, a series of strand invasions and branch migrations take place that frequently produce an intermediate consisting of two closely spaced Holliday junctions, often called a double Holliday junction (**Figure 5–64**).

Although some of the same proteins that function in double-strand break repair are used in meiosis, these proteins are directed by several meiosis-specific proteins to perform their tasks somewhat differently, resulting in the different DNA intermediates formed (compare Figure 5–59 with Figure 5–64). Another important difference is that, in meiosis, recombination occurs preferentially between maternal and paternal chromosomal homologs rather than between the newly replicated, identical DNA duplexes that pair in double-strand break repair.

There are two different ways to resolve the double Holliday intermediate shown in Figure 5–64. In the conceptually simplest resolution ("noncrossover"), the original pairs of crossing strands are cut at both Holliday junctions in the same way, which causes the two original helices to separate from one another in a form unaltered except for the region between the two junctions (see Figure 5–64, left; in that region, each helix contains a short region of

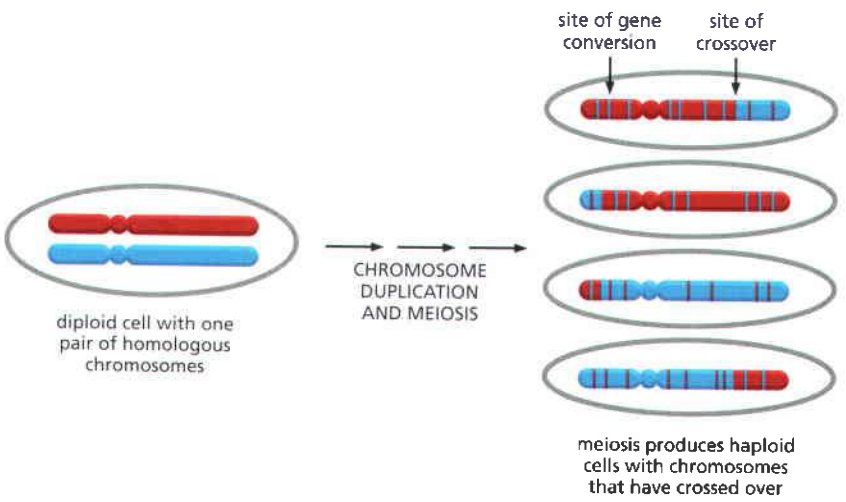


Figure 5–63 Chromosome crossing over occurs in meiosis. Meiosis is the process by which a diploid cell gives rise to four haploid germ cells, as described in detail in Chapter 21. Meiosis produces germ cells in which the paternal and maternal genetic information (*red* and *blue*) has been reassorted through chromosome crossovers. In addition, many short regions of gene conversion occur, as indicated.

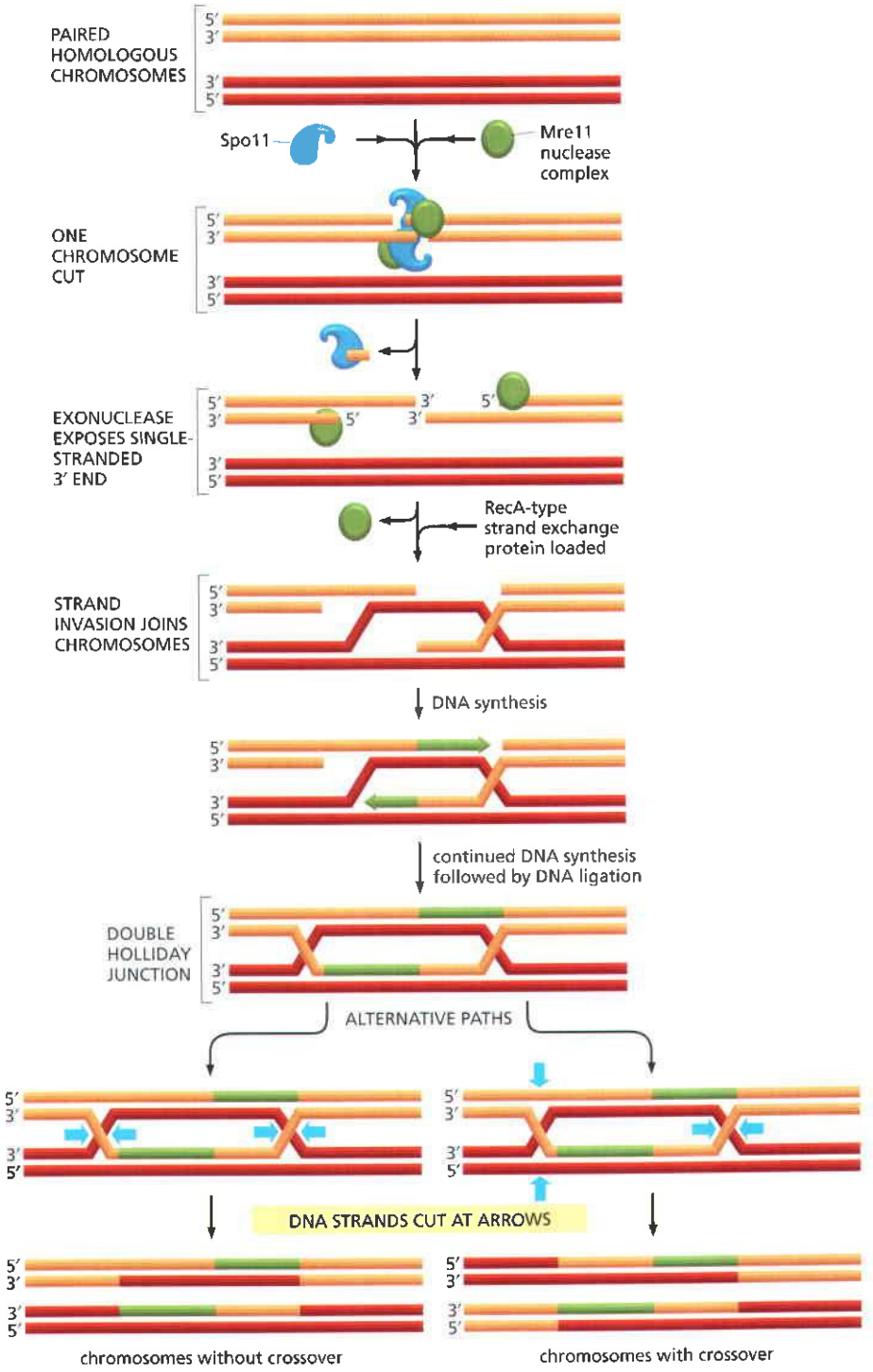


Figure 5–64 Homologous recombination in meiosis can generate crossovers. Once the meiosis-specific protein Spo11 and the Mre11 complex break the duplex DNA and process the ends, homologous recombination proceeds through a double Holliday junction. Many of the steps that produce chromosome crossovers in meiosis resemble those used to repair double-strand breaks (Figure 5–59). However, in meiosis, the process is tightly coupled to other meiotic events and it is directed by proteins, such as Spo11, that are only produced in meiotic cells.

heteroduplex adjacent to a region of homoduplex produced by DNA synthesis). If, on the other hand, the two Holliday junctions are resolved oppositely (one cleaved on the original pair of crossing strands and the other on the non-crossing strands), the outcome is much more profound. In this type of resolution ("crossover"), the portions of each chromosome upstream and downstream from the two Holliday junctions are swapped, creating two chromosomes that have crossed over (see Figure 5–64, right).

Relatively few of the Spo11-mediated double-strand breaks become crossovers; the majority (90% in humans, for example) are resolved as non-crossovers. It is not understood how this choice is made, but it apparently happens early in the recombination process, before the Holliday junctions are formed. The relatively few crossovers that do form are distributed along chromosomes such that the presence of a crossover in one position somehow inhibits crossing over in

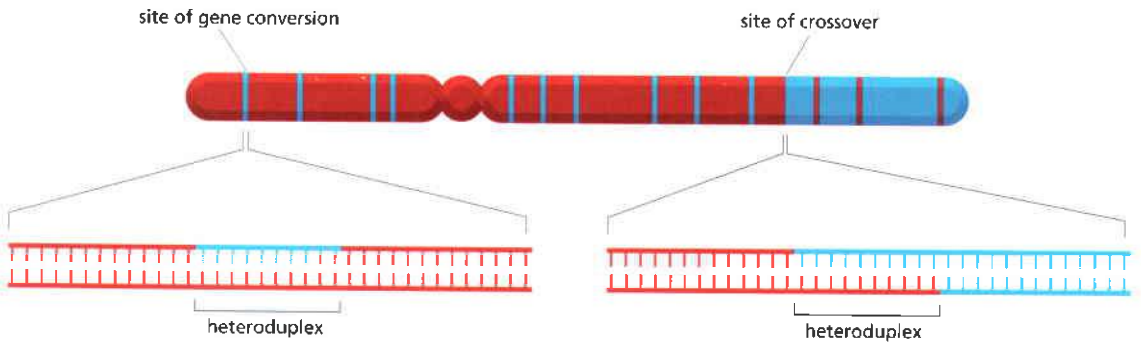


Figure 5–65 Heteroduplexes formed during meiosis. Heteroduplex DNA is present at sites of recombination that are resolved either as crossovers or non-crossovers.

neighboring regions. Termed *crossover control*, this fascinating but poorly understood regulatory mechanism presumably ensures the roughly even distribution of crossover points along chromosomes. For many organisms, roughly two crossovers per chromosome occur during each meiosis, one on each arm. As discussed in detail in Chapter 21, these crossovers play an important mechanical role in the proper segregation of chromosomes during meiosis.

Whether a meiotic recombination event is resolved as a crossover or a non-crossover, the recombination machinery leaves behind a *heteroduplex region* where a strand from the parental homolog is base-paired with a strand from the maternal homolog (**Figure 5–65**). These heteroduplex regions can tolerate a small percentage of mismatched base pairs, and they often extend for thousands of nucleotide pairs. Because of the many non-crossover events in meiosis, they produce scattered sites in the germ cells where short DNA sequences from one homolog have been pasted into the other homolog. And in all cases, they mark sites of potential *gene conversion*—that is, sites where the four haploid chromosomes produced by meiosis contain three copies of a short DNA sequence from one homolog and only one copy of this sequence from the other homolog (see Figure 5–63), as will now be explained.

Homologous Recombination Often Results in Gene Conversion

In sexually reproducing organisms, it is a fundamental law of genetics that each parent makes an equal genetic contribution to an offspring, which inherits one complete set of nuclear genes from the father and one complete set from the mother. Underlying this law is the highly accurate parceling out of chromosomes to the germ cells (eggs and sperm) that takes place during meiosis. Thus, when a diploid cell undergoes meiosis to produce four haploid germ cells (discussed in Chapter 21), exactly half of the genes distributed among these four cells should be maternal (genes that the diploid cell inherited from its mother) and the other half paternal (genes that the diploid cell inherited from its father) In some organisms (fungi, for example), it is possible to recover and analyze all four of the haploid gametes produced from a single cell by meiosis. Studies in such organisms have revealed rare cases in which the parceling out of genes violates the standard rules of genetics. Occasionally, for example, meiosis yields three copies of the maternal version of a gene and only one copy of the paternal allele (see Figure 5–63). Alternative versions of the same gene are called **alleles**, and the divergence from their expected distribution during meiosis is known as **gene conversion**. Genetic studies show that only small sections of DNA typically undergo gene conversion, and in many cases only a part of a gene is changed.

Several pathways in the cell can lead to gene conversion. First, the DNA synthesis that accompanies the early steps of homologous recombination will produce regions of the double Holliday junction where three copies of the sequence on one homolog are present (see *green* strands at the bottom of Figure 5–64); these will produce sites of gene conversion once the Holliday junction is resolved. In addition, if the two strands that make up a heteroduplex region do not have identical nucleotide sequences, mismatched pairs will result. These can be repaired by the cell's mismatch repair system, described earlier (see Figure 5–20). When used during recombination, however, the mismatch repair system makes

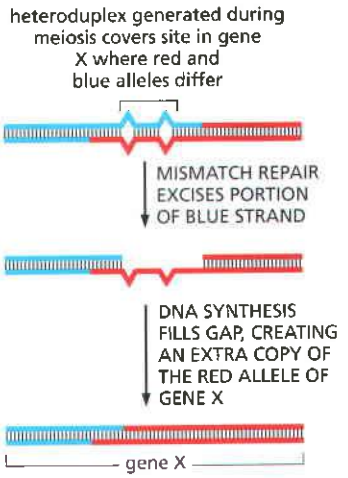


Figure 5–66 Gene conversion caused by mismatch correction. In this process, heteroduplex DNA is formed at the sites of homologous recombination between maternal and paternal chromosomes. If the maternal and paternal DNA sequences are slightly different, the heteroduplex region will include some mismatched base pairs, which may then be corrected by the DNA mismatch repair machinery (see Figure 5–20). Such repair can "erase" nucleotides on either the paternal or the maternal strand. The consequence of this mismatch repair is gene conversion, detected as a deviation from the segregation of equal copies of maternal and paternal alleles that normally occurs in meiosis.

no distinction between the paternal and maternal strand and will randomly choose which strand to repair. As a consequence of this repair, one allele will be "lost" and the other duplicated (Figure 5–66), resulting in net "conversion" of one allele to the other. Thus, gene conversion, originally regarded as a mysterious deviation from the rules of genetics, can be seen as a straightforward consequence of the mechanisms of homologous recombination and DNA repair.

Mismatch Proofreading Prevents Promiscuous Recombination Between Two Poorly Matched DNA Sequences

We have seen that homologous recombination relies on the pairing of complementary (or nearly complementary) DNA strands that initially come from separate DNA duplexes. But what controls how precise the matching must be? This is particularly crucial for recombination events that lead to crossovers. For example, the human genome contains many sets of closely related DNA sequences, and if crossing over were permitted between all of them, it would create havoc in the cell.

Although we do not completely understand how cells prevent inappropriate crossovers, we do know that components of the same mismatch proofreading system that removes replication errors (see Figure 5–20) and is responsible for some types of gene conversion (see Figure 5–66) have the additional role of interrupting genetic recombination between poorly matched DNA sequences. It is thought that the mismatch proofreading system normally recognizes the mispaired bases in an initial strand exchange, and—if there are significant mismatches—it prevents the subsequent steps (particularly branch migration) required to form a crossover. This type of recombinational proofreading is thought to prevent promiscuous recombination events that would otherwise scramble the human genome (Figure 5–67). Although controversial, it has also

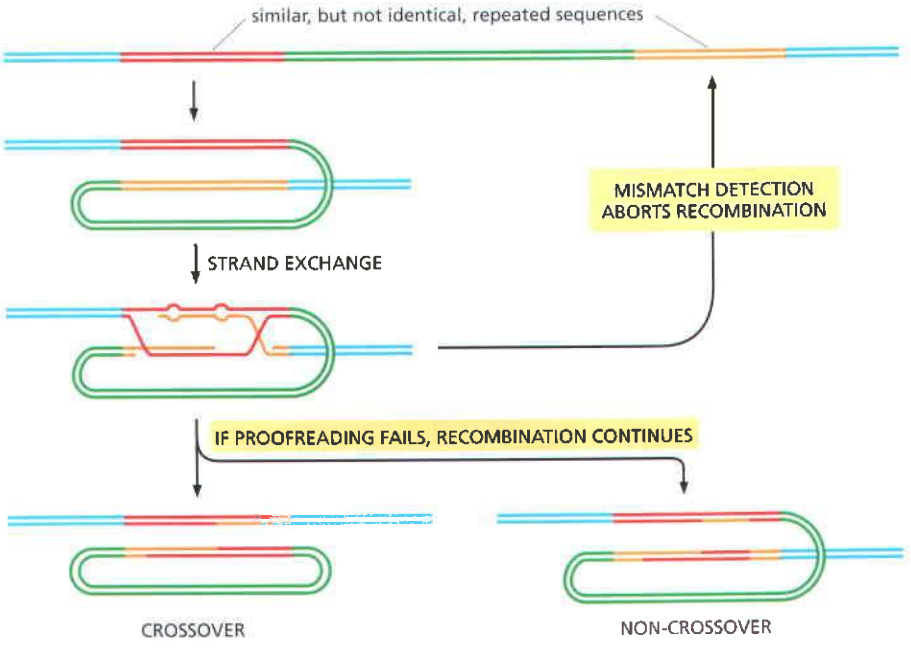


Figure 5–67 The mechanism that prevents general recombination from destabilizing a genome that contains repeated sequences. Components of the mismatch proofreading system, diagrammed in Figure 5–20, have the additional role of recognizing mismatches and preventing inappropriate recombination. If allowed to proceed, such recombination would produce deletions *(left)* or gene conversions *(right)* where information from one of the original repeated sequences has been lost.

been proposed that recombinational proofreading helps to preserve speciation, particularly among bacteria, by blocking genetic exchange between closely related species. For example, the genomes of *E. coli* and *Salmonella typhimurium* are 80% identical in nucleotide sequences, yet this proofreading step blocks recombination between their genomes.

Summary

Homologous recombination (also called general recombination) results in the transfer of genetic information between two DNA duplex segments of similar nucleotide sequence. This process is essential for the error-free repair of chromosome damage in all cells, and it is also responsible for the crossing-over of chromosomes that occurs during meiosis. The recombination event is guided by a specialized set of proteins. Although it can occur anywhere on a DNA molecule, it always requires extensive base-pairing interactions between complementary strands of the two interacting DNA duplexes.

In meiosis, homologous recombination is initiated by double-strand breaks that are intentionally produced along each chromosome. These breaks are processed into single-stranded 3′ ends which, in a reaction catalyzed by RecA-family proteins, invade a homologous partner DNA duplex. Branch migration accompanied by limited DNA synthesis then leads to the formation of four-stranded structures known as Holliday junctions. Each recombination reaction ends when DNA cutting resolves these recombination intermediates. The result can either be two chromosomes that have crossed-over (that is, chromosomes in which the DNA on either side of the site of DNA pairing originates from two different homologs), or two non-crossover chromosomes. In the latter case, the two chromosomes that result are identical to the original two homologs, except for relatively minor DNA sequence changes at the site of recombination. Unlike the situation in meiosis, the homologous recombination reactions that flawlessly repair double-strand breaks rarely produce crossover products.

TRANSPOSITION AND CONSERVATIVE SITE-SPECIFIC RECOMBINATION

We have seen that, through homologous recombination, rearrangements occur between DNA segments that can result in the exchange of DNA sequences between chromosomes. However, the order of genes on the interacting chromosome typically remains the same following homologous recombination, inasmuch as the recombining sequences must be very similar for the process to occur. In this section, we describe two very different types of recombination—**transposition** (also called *transpositional recombination*) and **conservative site-specific recombination**—that do not require substantial regions of DNA homology. These two types of recombination events can alter gene order along a chromosome, and cause unusual types of mutations that add new information to genomes.

Transposition and conservative site-specific recombination are largely dedicated to moving a wide variety of specialized segments of DNA, collectively termed *mobile genetic elements*, from one position in a genome to another. We will see that mobile genetic elements can range in size from a few hundred to tens of thousands of nucleotide pairs, and each typically carries a unique set of genes. Often, one of these genes encodes a specialized enzyme that catalyzes the movement of only that element, thereby making this type of recombination possible.

Virtually all cells contain mobile genetic elements (known informally as "jumping genes"). As explained in Chapter 4, over evolutionary timescales, they have had a profound effect on the shaping of modern genomes. For example, nearly half of the human genome can be traced to these elements (see Figure 4–17). Over time, random mutation has altered their nucleotide sequences, and, as a result, only a few of the many copies of these elements in our DNA are still active and capable of movement. The remainder are molecular fossils whose existence provides striking clues about our own evolutionary history.

Mobile genetic elements are often considered to be molecular parasites (they are also termed "selfish DNA") that persist because cells cannot get rid of them; they certainly have come close to overrunning our own genome. However, mobile DNA elements can provide benefits to the cell. For example, the genes they carry are sometimes advantageous, as in the case of antibiotic resistance in bacterial cells discussed below. The movement of mobile genetic elements also produces many of the genetic variants upon which evolution depends, because, in addition to moving themselves, mobile genetic elements occasionally rearrange neighboring sequences of the host genome. Thus, spontaneous mutations observed in *Drosophila*, humans, and other organisms are often due to the movement of mobile genetic elements. While the vast majority of these mutations will be deleterious to the organism, some will result in increased fitness and tend to spread throughout the population. It is almost certain that much of the variety of life we see around us originally arose from the movement of mobile genetic elements.

In this section, we introduce mobile genetic elements and describe the mechanisms that enable them to move around a genome. We shall see that some of these elements move through transposition mechanisms and others through conservative site-specific recombination. We begin with transposition, as there are many more examples of this type of movement known.

Through Transposition, Mobile Genetic Elements Can Insert Into Any DNA Sequence

Mobile elements that move by way of transposition are called **transposons**, or **transposable elements**. In transposition, a specific enzyme, usually encoded by the transposon itself and typically called a *transposase*, acts on a specific DNA sequence at each end of the transposon, causing it to insert into a new target DNA site. Most transposons are only modestly selective in choosing their target site, and they can therefore insert themselves into many different locations in the genome. In particular, there is no general requirement for homology between the ends of the element and the target sequence. Most transposons move only rarely. In bacteria, where it is possible to measure the frequency accurately, transposons typically move once every 10^5 cell divisions. In most cases, transposition appears to be a rare stochastic process, albeit one often linked to the passage of a replication fork.

On the basis of their structure and transposition mechanism, transposons can be grouped into three large classes: *DNA-only transposons, retroviral-like retrotransposons,* and *nonretroviral retrotransposons*. Each class will be discussed in detail below. For reference purposes, the differences between them are briefly outlined in **Table 5–3**.

DNA-Only Transposons Move by Both Cut-and-Paste and Replicative Mechanisms

DNA-only transposons predominate in bacteria, and they are largely responsible for the spread of antibiotic resistance in bacterial strains. When antibiotics like penicillin and streptomycin first became widely available in the 1950s, most bacteria that caused human disease were susceptible to them. Fifty years later, the situation has changed dramatically—antibiotics such as penicillin (and its modern derivatives) are no longer effective against many modern bacterial strains, including those causing gonorrhea and bacterial pneumonia. The spread of antibiotic resistance is due largely to genes that encode antibiotic-inactivating enzymes that are carried on transposons (**Figure 5–68**). Although these mobile elements can transpose only within cells that already carry them, they can be moved from one cell to another through other mechanisms known collectively as horizontal gene transfer (Figure 1–23). Once introduced into a new cell, a transposon can insert itself into the genome and be faithfully passed on to all its progeny cells through the normal processes of DNA replication and cell division.

Table 5–3 Three Major Classes of Transposable Elements

CLASS DESCRIPTION AND STRUCTURE	SPECIALIZED ENZYMES REQUIRED FOR MOVEMENT	MODE OF MOVEMENT	EXAMPLES
DNA-only transposons			
short inverted repeats at each end	transposase	moves as DNA, either by cut-and-paste or replicative pathways	P element (*Drosophila*) Ac-Ds (maize) Tn3 and Tn10 (*E. coli*) Tam3 (snapdragon)
Retroviral-like retrotransposons			
directly repeated long terminal repeats (LTRs) at each end	reverse transcriptase and integrase	moves via an RNA intermediate produced by a promoter in the LTR	Copia (*Drosophila*) Ty1 (yeast) THE1 (human) Bs1 (maize)
Nonretroviral retrotransposons			
Poly A at 3′ end of RNA transcript; 5′ end is often truncated	reverse transcriptase and endonuclease	moves via an RNA intermediate that is often produced from a neighboring promoter	F element (*Drosophila*) L1 (human) Cin4 (maize)

These elements range in length from 1000 to about 12,000 nucleotide pairs. Each family contains many members, only a few of which are listed here. In addition to transposable elements, some viruses can move in and out of host cell chromosomes by transpositional mechanisms. These viruses are related to the first two classes of transposons.

DNA-only transposons, so named because they exist only as DNA during their movement, can relocate from a donor site to a target site either by *cut-and-paste transposition* or by *replicative transposition*. Because it is conceptually simpler, we discuss the cut-and-paste mechanism first. The process begins when each of the special short DNA sequences that mark the two ends of the element binds a molecule of transposase. The two transposase molecules come together to form a multimeric "transpososome" that produces a DNA loop juxtaposing the two ends of the element (**Figure 5–69**). The transposase then introduces cuts at the base of the loop and removes the element completely from its original chromosome, forming the central intermediate in the transposition process (**Figure 5–70**). To complete the DNA movement, the transposase catalyzes a direct attack of the element's two DNA ends on a target DNA molecule, breaking two phosphodiester bonds in the target molecule and creating two new ones as it joins the element and target DNAs together. Because this DNA joining reaction begins and ends with the same number of phosphodiester bonds, it can occur without the input of additional energy. We will see in the next chapter that this same type of phosphodiester bond rearrangement (called *transesterification*) underlies another fundamental process in molecular biology, RNA splicing.

Because the breaks made in the two target DNA strands are staggered (*red arrowheads* in Figure 5–69), the product DNA molecule initially contains two

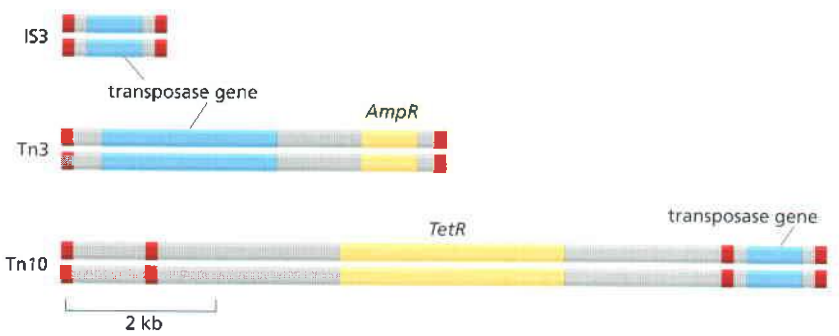


Figure 5–68 Three of the many DNA-only transposons found in bacteria. Each of these mobile DNA elements contains a gene that encodes a *transposase*, an enzyme that conducts at least some of the DNA breakage and joining reactions needed for the element to move. Each transposon also carries short DNA sequences (indicated in *red*) that are recognized only by the transposase encoded by that element and are necessary for movement of the element. In addition, two of the three mobile elements shown carry genes that encode enzymes that inactivate the antibiotics ampicillin (*AmpR*) and tetracycline (*TetR*). The transposable element Tn10, shown in the bottom diagram, is thought to have evolved from the chance landing of two much shorter mobile elements on either side of a tetracycline-resistance gene: the wide use of tetracycline as an antibiotic has selected for the spread of this transposon through bacterial populations.

short, single-stranded gaps, one at each end of the inserted transposon. A host-cell DNA polymerase and DNA ligase fill in and seal these gaps to complete the recombination process. This produces a short duplication of the target DNA sequence at the insertion site; these flanking direct repeat sequences, whose length is different for different transposons, serve as convenient records of prior transposition events.

When a cut-and-paste DNA-only transposon is excised from its original location, it leaves behind a "hole" in the chromosome. This lesion can be perfectly healed by recombinational double-strand break repair (see Figure 5–59), provided that the chromosome has just been replicated and an identical copy of the damaged host sequence is available. In this case, the repair process will restore the transposon to its original position. Alternatively, in diploid organisms, the damaged chromosome can be recombinationally repaired using the chromosomal homolog, in which case the transposon will not be restored but a loss of heterozygosity could occur at the site of repair. As the third possibility, a nonhomologous end-joining reaction can reseal the break; in this case, the DNA sequence that originally flanked the transposon is altered, producing a mutation at the chromosomal site from which the transposon was excised (see Figure 5–52).

Remarkably, the same mechanism used to excise cut-and-paste transposons from DNA has been found to operate in developing immune systems of vertebrates, catalyzing the DNA rearrangments that produce antibody and T cell receptor diversity. Known as *V(D)J recombination*, this process will be discussed in Chapter 25. Found only in vertebrates, V(D)J recombination is a relatively recent evolutionary novelty, but it is believed to have evolved from the much more ancient cut-and-paste transposons.

Some DNA-only transposons move by a mechanism called *replicative transposition*. In this case, the transposon DNA is replicated during transposition: one copy remains at the original site while the other is inserted at a new chromosomal location. Although the mechanism used is more complex, it is closely related to the cut-and-paste mechanism just described; indeed, some transposons can move by both pathways.

Some Viruses Use a Transposition Mechanism to Move Themselves into Host Cell Chromosomes

Certain viruses are considered mobile genetic elements because they use transposition mechanisms to integrate their genomes into that of their host cell. However, unlike transposons, these viruses encode proteins that package their genetic information into virus particles that can infect other cells. Many of the

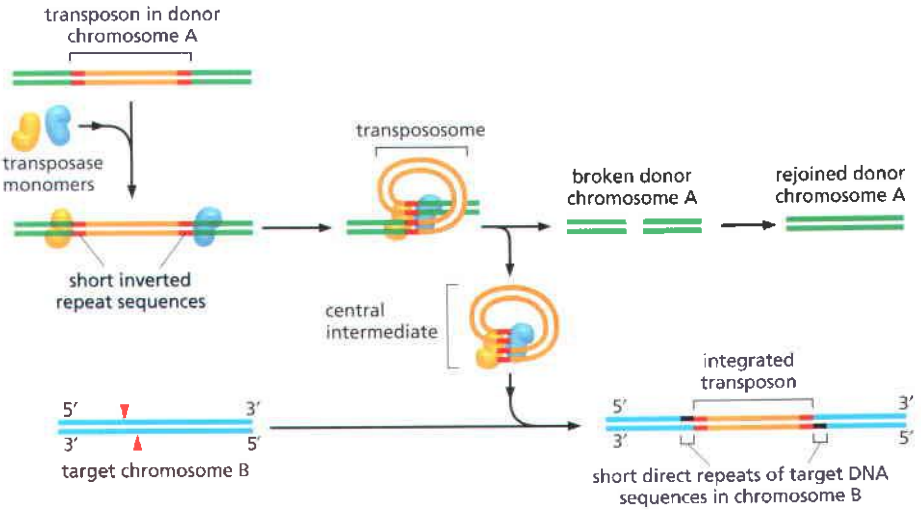


Figure 5–69 Cut-and-paste transposition. DNA-only transposons can be recognized in chromosomes by the "inverted repeat DNA sequences" (*red*) present at their ends. These sequences, which can be as short as 20 nucleotides, are all that is necessary for the DNA between them to be transposed by the particular transposase enzyme associated with the element. The cut-and-paste movement of a DNA-only transposable element from one chromosomal site to another begins when the transposase brings the two inverted DNA sequences together, forming a DNA loop. Insertion into the target chromosome, catalyzed by the transposase, occurs at a random site through the creation of staggered breaks in the target chromosome (*red arrowheads*). Following the transposition reaction, the single-strand gaps created by the staggered break are repaired by DNA polymerase and ligase (*purple*). As a result, the insertion site is marked by a short direct repeat of the target DNA sequence, as shown. Although the break in the donor chromosome (*green*) is repaired, this process often alters the DNA sequence, causing a mutation at the original site of the excised transposable element (not shown).

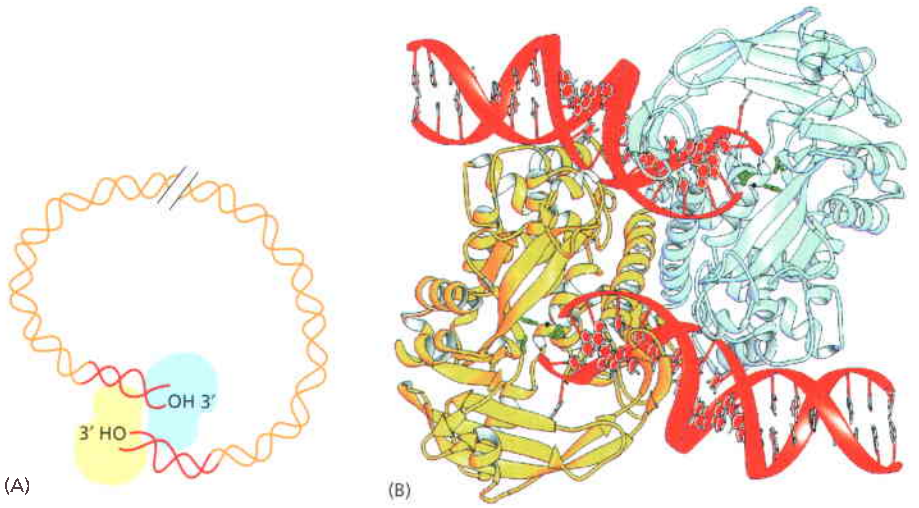


Figure 5–70 The structure of the central intermediate formed by a cut-and-paste transposase. (A) Schematic view of the overall structure. (B) The detailed structure of a transposase holding the two DNA ends, whose 3′-OH groups are poised to attack a target chromosome. One domain of the transposase recognizes the DNA sequence at the transposon end while a different domain carries out the DNA breakage–joining chemistry (B, from D.R. Davies et al., *Science* 289:77–85, 2000. With permission from AAAS.)

viruses that insert themselves into a host chromosome do so by employing one of the top two mechanisms listed in Table 5–3. Indeed, much of our knowledge of these mechanisms has come from studies of particular viruses that employ them.

A virus that infects a bacterium is known as a **bacteriophage**. The *bacteriophage Mu* not only uses DNA-based transposition to integrate its genome into its host cell chromosome, it also uses replicative transposition to replicate its genome. Transposition also has a key role in the life cycle of many other viruses. Most notable are the **retroviruses**, which include the human AIDS virus, HIV. Outside the cell, a retrovirus exists as a single-stranded RNA genome packed into a protein capsid along with a virus-encoded **reverse transcriptase** enzyme. During the infection process, the viral RNA enters a cell and is converted to a double-stranded DNA molecule by the action of this crucial enzyme, which is able to polymerize DNA on either an RNA or a DNA template (**Figure 5–71** and **Figure 5–72**). The term *retrovirus* refers to the virus's ability to reverse the usual flow of genetic information, which normally is from DNA to RNA (see Figure 1–5).

Once the reverse transcriptase has produced a double-stranded DNA molecule, specific sequences near its two ends are held together by a virus-encoded transposase called *integrase*. Integrase creates activated 3′-OH viral DNA ends that can directly attack a target DNA molecule through a mechanism similar to that used by the cut-and-paste DNA-only transposons (**Figure 5–73**). In fact, detailed analyses of the three-dimensional structures of bacterial transposases and HIV integrase have revealed remarkable similarities in these enzymes, even though their amino acid sequences have diverged considerably.

Retroviral-like Retrotransposons Resemble Retroviruses, but Lack a Protein Coat

A large family of transposons called **retroviral-like retrotransposons** (see Table 5–3) move themselves in and out of chromosomes by a mechanism that is identical to that used by retroviruses. These elements are present in organisms as diverse as yeasts, flies, and mammals; unlike viruses, they have no intrinsic ability to leave their resident cell but are passed along to all descendants of that cell through the normal process of DNA replication and cell division. The first step in their transposition is the transcription of the entire transposon, producing an RNA copy of the element that is typically several thousand nucleotides long. This transcript, which is translated as a messenger RNA by the host cell, encodes a reverse transcriptase enzyme. This enzyme makes a double-stranded DNA copy of the RNA molecule via an RNA/DNA hybrid intermediate, precisely mirroring the early stages of infection by a retrovirus (see Figure 5–71). Like retroviruses, the linear double-stranded DNA molecule then integrates into a site on the chromosome by using an integrase enzyme that is also encoded by the element (see Figure 5–73).

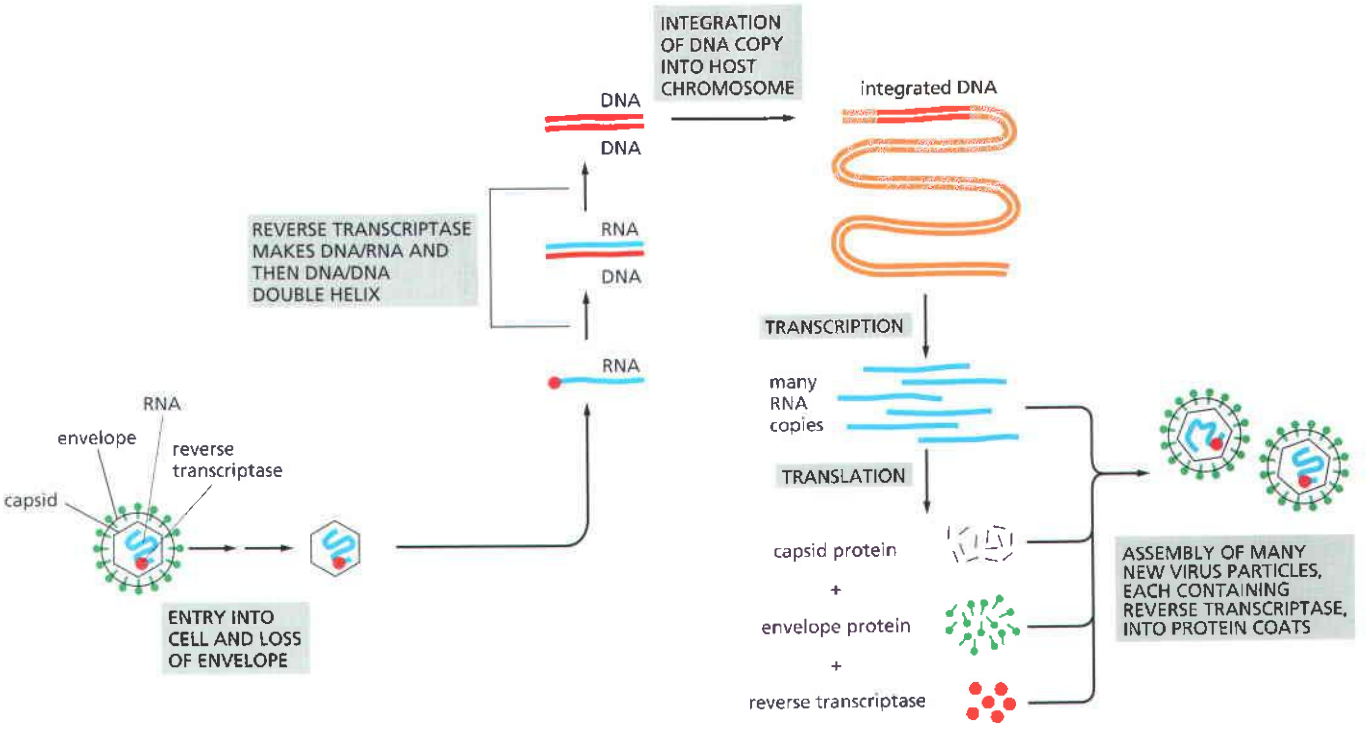
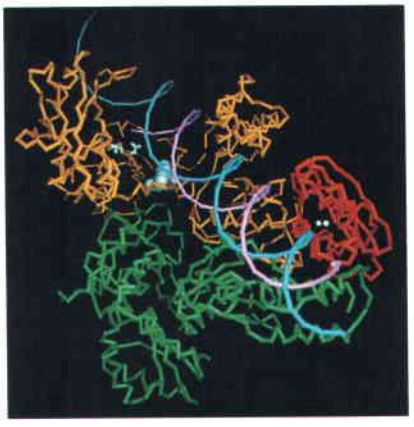


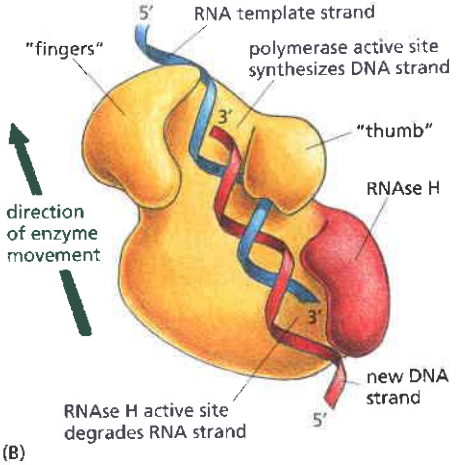
Figure 5–71 The life cycle of a retrovirus. The retrovirus genome consists of an RNA molecule of about 8500 nucleotides; two such molecules are typically packaged into each viral particle. The enzyme reverse transcriptase first makes a DNA copy of the viral RNA molecule and then a second DNA strand, generating a double-stranded DNA copy of the RNA genome. The integration of this DNA double helix into the host chromosome is then catalyzed by a virus-encoded integrase enzyme (see Figure 5–73). This integration is required for the synthesis of new viral RNA molecules by the host cell RNA polymerase, the enzyme that transcribes DNA into RNA (discussed in Chapter 6).

A Large Fraction of the Human Genome Is Composed of Nonretroviral Retrotransposons

A significant fraction of many vertebrate chromosomes is made up of repeated DNA sequences. In human chromosomes, these repeats are mostly mutated and truncated versions of **nonretroviral retrotransposons**, the third major type of transposon (see Table 5–3). Although most of these transposons are immobile, a few retain the ability to move. Relatively recent movements of the *L1 element* (sometimes referred to as a LINE or long interspersed nuclear element) have been identified, some of which result in human disease; for example, a particular type of hemophilia results from an *L1* insertion into the gene encoding the blood clotting protein Factor VIII (see Figure 6–25).



(A)



(B)

Figure 5–72 Reverse transcriptase. (A) The three-dimensional structure of the enzyme from HIV (the human AIDS virus) determined by X-ray crystallography. (B) A model showing the enzyme's activity on an RNA template. Note that the polymerase domain *(yellow* in B) has a covalently attached RNAse H (H for "hybrid") domain *(red)* that degrades the RNA strand in an RNA/DNA helix. This activity helps the polymerase to convert the initial hybrid helix into a DNA double helix (A, courtesy of Tom Steitz; B, adapted from L.A. Kohlstaedt et al., *Science* 256:1783–1790, 1992. With permission from AAAS.)

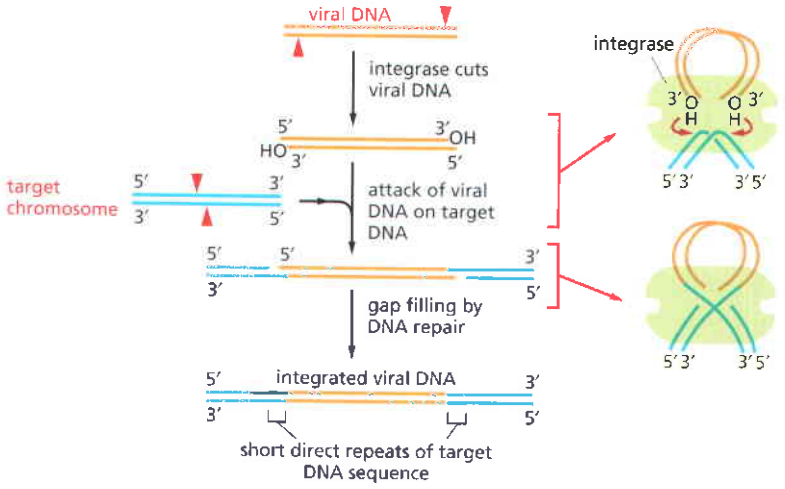


Figure 5–73 Transposition by either a retrovirus (such as HIV) or a retroviral-like retrotransposon. The process begins with a double-stranded DNA molecule *(orange)* produced by reverse transcriptase (see Figure 5–71). In an initial step, the integrase enzyme forms a DNA loop and cuts one strand at each end of the viral DNA sequence, exposing new 3′-OH groups. Each of these 3′-OH ends then directly attack a phosphodiester bond on opposite strands of a randomly selected site on a target chromosome *(red arrowheads* on *blue* DNA). This inserts the viral DNA sequence into the target chromosome, leaving short gaps on each side that are filled in by DNA repair processes. Because of the gap filling, this mechanism (like that of cut-and-paste transposition) leaves short repeats of target DNA sequence *(black)* on each side of the integrated DNA segment.

Nonretroviral retrotransposons are found in many organisms and move via a distinct mechanism that requires a complex of an endonuclease and a reverse transcriptase. As illustrated in **Figure 5–74**, the RNA and reverse transcriptase have a much more direct role in the recombination event than they do in the retroviral-like retrotransposons described above.

Inspection of the human genome sequence has revealed that the bulk of nonretroviral retrotransposons—for example, the many copies of the *Alu* element, a member of the SINE (short interspersed nuclear element) family—do not carry their own endonuclease or reverse transcriptase genes. Nonetheless they have successfully amplified themselves to become major constituents of our genome, presumably by pirating enzymes encoded by other transposons.

The *L1* and *Alu* elements seem to have multiplied in the human genome relatively recently (**Figure 5–75**). Thus, for example, the mouse contains sequences closely related to *L1* and *Alu*, but their placement in mouse chromosomes differs from that in human chromosomes (see Figure 4–80). Together the LINEs and SINEs make up about 40% of the human genome (see Figure 4–17).

Different Transposable Elements Predominate in Different Organisms

We have described several types of transposable elements: (1) DNA-only transposons, the movement of which is based on DNA breaking and joining reactions; (2) retroviral-like retrotransposons, which also move via DNA breakage and joining, but where RNA has a key role as a template to generate the DNA recombination substrate; and (3) nonretroviral retrotransposons, in which an RNA copy of the element is central to the incorporation of the element into the target DNA, acting as a direct template for a DNA target-primed reverse transcription event.

Intriguingly, different types of transposons predominate in different organisms. For example, the vast majority of bacterial transposons are DNA-only types, with a few related to the nonretroviral retrotransposons also present. In yeasts, the main mobile elements that have been observed are retroviral-like retrotransposons. In *Drosophila*, DNA-based, retroviral, and nonretroviral

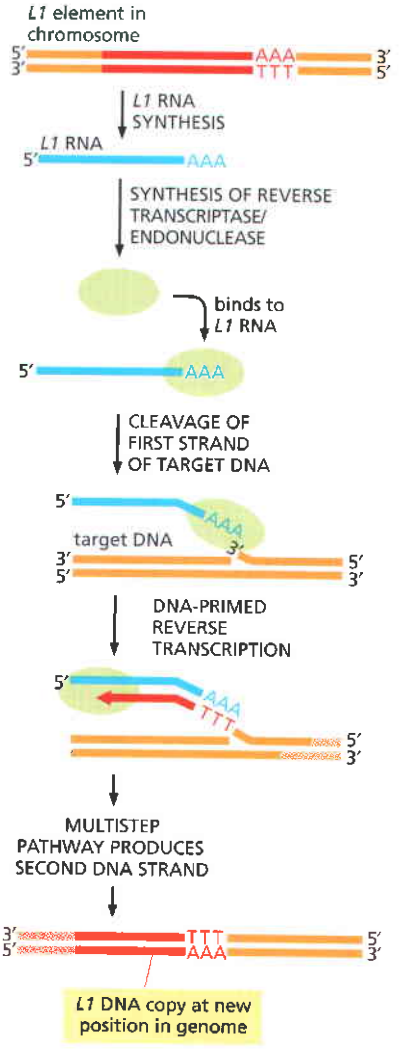


Figure 5–74 Transposition by a nonretroviral retrotransposon. Transposition by the *L1* element *(red)* begins when an endonuclease attached to the *L1* reverse transcriptase *(green)* and the *L1* RNA *(blue)* nicks the target DNA at the point at which insertion will occur. This cleavage releases a 3′-OH DNA end in the target DNA, which is then used as a primer for the reverse transcription step shown. This generates a single-stranded DNA copy of the element that is directly linked to the target DNA. In subsequent reactions, further processing of the single-stranded DNA copy results in the generation of a new double-stranded DNA copy of the *L1* element that is inserted at the site of the initial nick.

transposons are all found. Finally, the human genome contains all three types of transposon, but as discussed below, their evolutionary histories are strikingly different.

Genome Sequences Reveal the Approximate Times that Transposable Elements Have Moved

The nucleotide sequence of the human genome provides a rich "fossil record" of the activity of transposons over evolutionary time spans. By carefully comparing the nucleotide sequences of the approximately 3 million transposable element remnants in the human genome, it has been possible to broadly reconstruct the movements of transposons in our ancestors' genomes over the past several hundred million years. For example, the DNA-only transposons appear to have been very active well before the divergence of humans and Old World monkeys (25–35 million years ago); but, because they gradually accumulated inactivating mutations, they have been dormant in the human lineage since that time. Likewise, although our genome is littered with relics of retroviral-like transposons, none appear to be active today. Only a single family of retroviral-like retrotransposons is believed to have transposed in the human genome since the divergence of human and chimpanzee approximately 6 million years ago. The nonretroviral retrotransposons are also ancient, but in contrast to other types, some are still moving in our genome, as mentioned previously. For example, it is estimated that *de novo* movement of an *Alu* element is seen once every 100–200 human births. The movement of nonretroviral retrotransposons is responsible for a small but significant fraction of new human mutations—perhaps two mutations out of every thousand.

The situation in mice is significantly different. Although the mouse and human genomes contain roughly the same density of the three types of transposons, both types of retrotransposons are still actively transposing in the mouse genome, being responsible for approximately 10% of new mutations.

Although we are only beginning to understand how the movements of transposons has shaped the genomes of present-day mammals, it has been proposed that bursts in transposition activity could have been responsible for critical speciation events during the radiation of the mammalian lineages from a common ancestor, a process that began approximately 170 million years ago. At this point, we can only wonder how many of our uniquely human qualities arose from the past activity of the many mobile genetic elements whose remnants are found today scattered throughout our chromosomes.

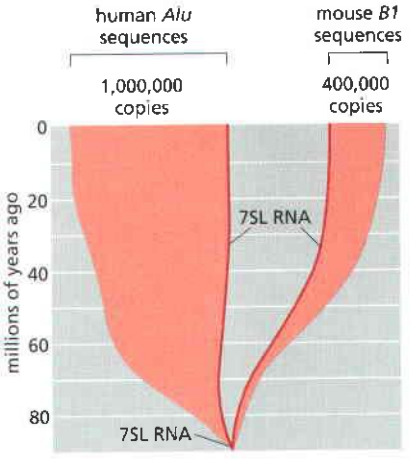


Figure 5–75 The expansion of the abundant *Alu* and *B1* sequences found in the human and mouse genomes, respectively. Both of these transposable DNA sequences are thought to have evolved from the 7SL RNA gene, which encodes the SRP RNA (see Figure 12–39). On the basis of their positions in the two genomes and the sequence similarity of these highly repeated elements, the major expansions in copy number seem to have occurred independently in mice and humans (see Figure 4–80). (Adapted from P.L. Deininger and G.R. Daniels, *Trends Genet.* 2:76–80, 1986, with permission from Elsevier, and International Human Genome Sequencing Consortium, *Nature* 409:860–921, 2001, with permission from Macmillan Publishers Ltd.)

Conservative Site-specific Recombination Can Reversibly Rearrange DNA

A different kind recombination mechanism, known as *conservative site-specific recombination,* mediates the rearrangements of other types of mobile DNA elements. In this pathway, breakage and joining occur at two special sites, one on each participating DNA molecule. Depending on the positions and relative orientations of the two recombination sites, DNA integration, DNA excision, or DNA inversion can occur (**Figure 5–76**). Conservative site-specific recombination is carried out by specialized enzymes that break and rejoin two DNA double helices at specific sequences on each DNA molecule. The same enzyme system that joins two DNA molecules can often take them apart again, precisely restoring the sequence of the two original DNA molecules (see Figure 5–76A).

Several features distinguish conservative site-specific recombination from transposition. First, conservative site-specific recombination requires specialized DNA sequences on both the donor and recipient DNA (hence the term site-specific). These sequences contain recognition sites for the particular recombinase that will catalyze the rearrangement. In contrast, transposition requires only that the transposon have a specialized sequence; for most transposons, the recipient DNA can be of any sequence. Second, the reaction mechanisms are

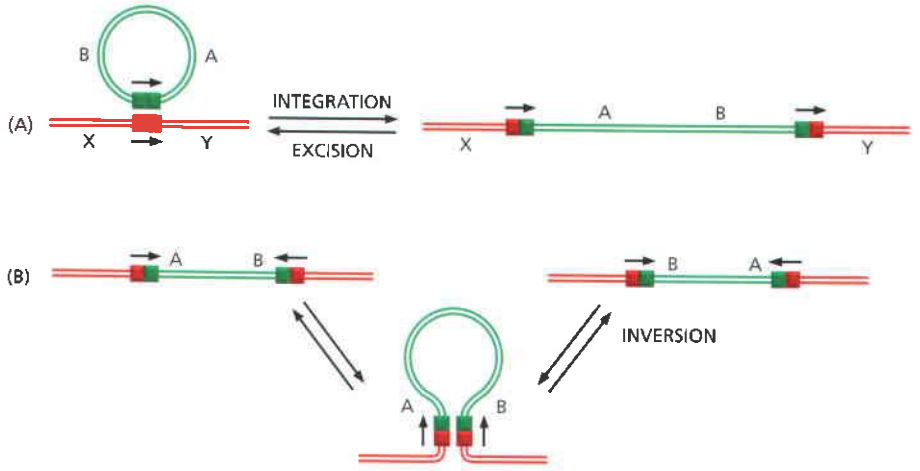


Figure 5–76 Two types of DNA rearrangement produced by conservative site-specific recombination. The only difference between the reactions in (A) and (B) is the relative orientation of the two short DNA sites (indicated by *arrows*) at which a site-specific recombination event occurs. (A) Through an integration reaction, a circular DNA molecule can become incorporated into a second DNA molecule; by the reverse reaction (excision), it can exit to reform the original DNA circle. Many bacterial viruses move in and out of their host chromosomes in this way (see Figure 5–77). (B) Conservative site-specific recombination can also invert a specific segment of DNA in a chromosome. A well-studied example of DNA inversion through site-specific recombination occurs in the bacterium *Salmonella typhimurium*, an organism that is a major cause of food poisoning in humans; the inversion of a DNA segment changes the type of flagellum that is produced by the bacterium (see Figure 7–64).

fundamentally different. The recombinases that catalyze conservative site-specific recombination resemble topoisomerases in the sense that they form transient high-energy covalent bonds with the DNA and use this energy to complete the DNA rearrangements. Thus all the phosphate bonds that are broken during a recombination event are restored upon its completion (hence the term conservative). Transposition, in contrast, uses a transesterification reaction that does not proceed through a covalently joined protein–DNA intermediate. This process leaves gaps in the DNA that must be resealed by DNA polymerase and ligase, both of which require the input of energy in the form nucleotide hydrolysis.

Conservative Site-Specific Recombination Was Discovered in Bacteriophage λ

A bacterial virus, *bacteriophage lambda*, was the first mobile DNA element of any type to be understood in biochemical detail. When this virus enters a cell, it directs the synthesis of a virus-encoded recombinase enzyme called *lambda integrase*. This enzyme mediates the covalent joining of the viral DNA to the bacterial chromosome, causing the virus to become part of this chromosome so that it is replicated automatically—as part of the host's DNA. The recombination process begins when several molecules of the integrase protein bind tightly to a specific DNA sequence on the circular bacteriophage chromosome, along with several host proteins. This DNA–protein complex can now bind to an attachment-site DNA sequence on the bacterial chromosome, bringing the bacterial and bacteriophage chromosomes together. The integrase then catalyzes the required cutting and resealing reactions that result in recombination. Because of a short region of sequence homology in the two joined sequences, a tiny heteroduplex joint is formed at this point of exchange (**Figure 5–77**).

The same type of site-specific recombination mechanism enables bacteriophage lambda DNA to exit from its integration site in the *E. coli* chromosome in response to specific signals and multiply rapidly within the bacterial cell (**Figure 5–78**). Excision is catalyzed by a complex of integrase enzyme and host factors with a second bacteriophage protein, excisionase, which is produced by the virus only when its host cell is stressed—in which case, it is in the bacteriophage's interest to abandon the host cell and multiply again as a virus particle.

Conservative Site-Specific Recombination Can Be Used to Turn Genes On or Off

When the special sites recognized by a conservative site-specific recombination enzyme are inverted in their orientation, the DNA sequence between them is inverted rather than excised (see Figure 5–76). Many bacteria use such an inversion of a DNA sequence to control the gene expression of particular genes—for example, by assembling an active gene from separated coding segments. This

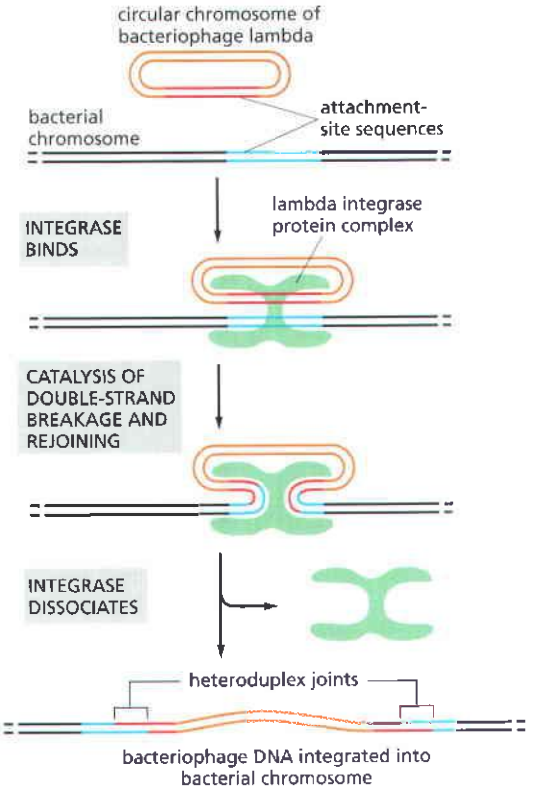


Figure 5–77 The insertion of a circular bacteriophage lambda DNA chromosome into the bacterial chromosome. In this example of conservative site-specific recombination, the lambda integrase enzyme binds to a specific attachment-site DNA sequence on each chromosome, where it makes cuts that bracket a short homologous DNA sequence. The integrase then switches the partner strands and reseals them to form a heteroduplex joint that is seven nucleotide pairs long. A total of four strand-breaking and strand-joining reactions are required; for each of them, the energy of the cleaved phosphodiester bond is stored in a transient covalent linkage between the DNA and the enzyme, so that DNA strand resealing occurs without a requirement for ATP or DNA ligase.

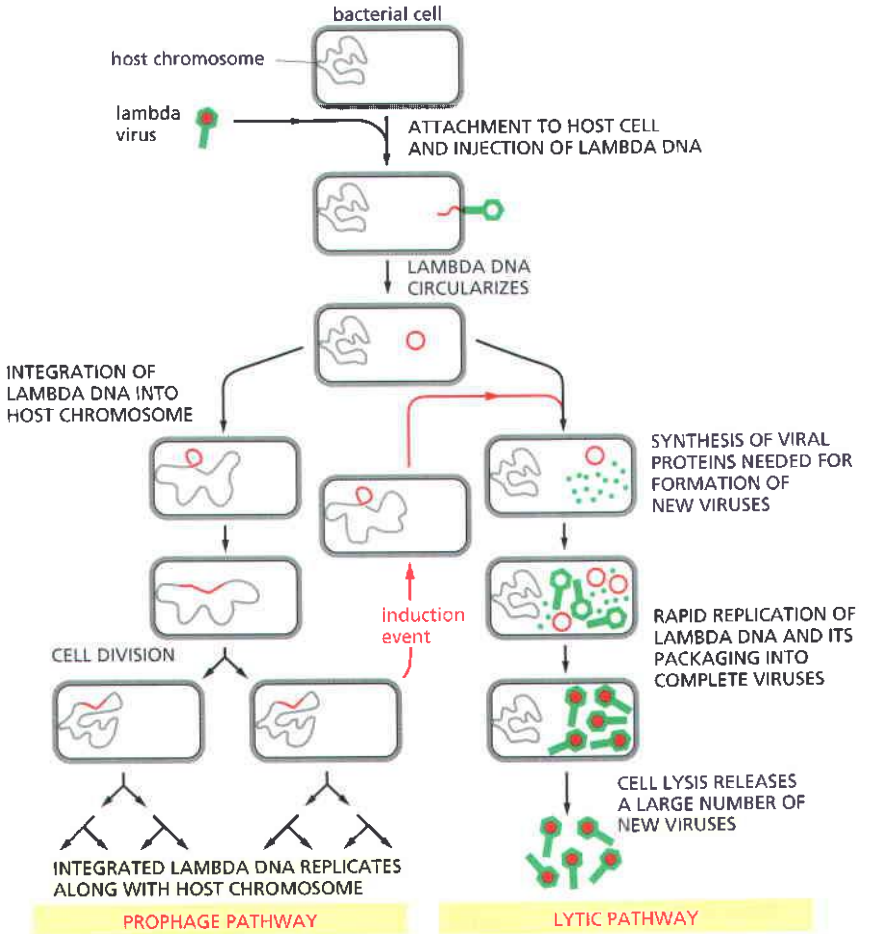


Figure 5–78 The life cycle of bacteriophage lambda. The double-stranded DNA lambda genome contains 50,000 nucleotide pairs and encodes 50–60 different proteins. When the lambda DNA enters the cell, its two ends join to form a circular DNA molecule. This bacteriophage can multiply in *E. coli* by a lytic pathway, which destroys the cell, or it can enter a latent prophage state. Damage to a cell carrying a lambda prophage induces the prophage to exit from the host chromosome and shift to lytic growth *(red arrows)*. Both the entrance of the lambda DNA into, and its exit from, the bacterial chromosome are accomplished by conservative site-specific recombination catalyzed by the lambda integrase enzyme (see Figure 5–77).

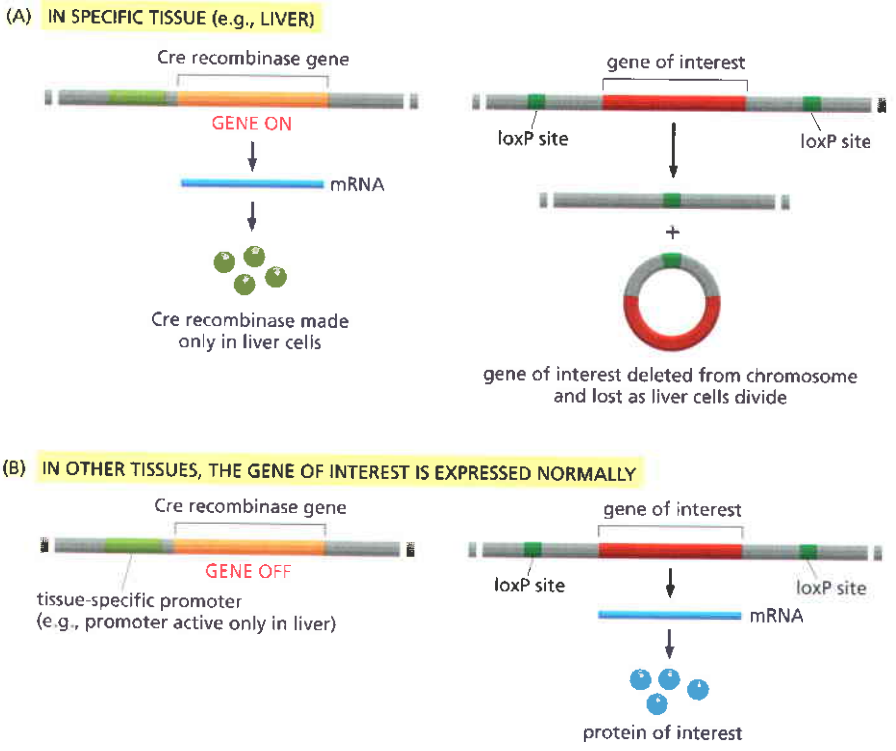


Figure 5–79 How a conservative site-specific recombination enzyme from bacteria can be used to delete specific genes from particular mouse tissues. This approach requires the insertion of two specially engineered DNA molecules into the animal's germ line. The first contains the gene for a recombinase (in this case the Cre recombinase from the bacteriophage P1) under the control of a tissue-specific promoter, which ensures that the recombinase is expressed only in that tissue. The second DNA molecule contains the gene of interest flanked by recognition sites (in this case loxP sites) for the recombinase. The mouse is engineered so that this is the only copy of this gene. Therefore, if the recombinase is expressed only in the liver, the gene of interest will be deleted there, and only there.

As described in Chapter 7, many tissue-specific promoters are known; moreover, many of these promoters are active only at specific times in development. Thus, it is possible to study the effect of deleting specific genes at many different times during the development of each tissue.

type of gene control has the advantage of being directly inheritable, since the new DNA arrangement is transferred to daughter chromosomes automatically when a cell divides. We will encounter a specific example of this use of conservative site-specific recombination in Chapter 7 (see Figure 7–64).

Bacterial conservative site-specific recombinases have also become powerful tools for cell and developmental biologists. To decipher the roles of specific genes and proteins in complex multicellular organisms, genetic engineering techniques are used to produce mice carrying a gene encoding a site-specific recombination enzyme plus a carefully designed target DNA with the DNA sites that this enzyme recognizes. At an appropriate time, the gene encoding the enzyme can be activated to rearrange the target DNA sequence. Such a rearrangement is widely used to delete a specific gene in a particular tissue of the mouse (**Figure 5–79**). It is particularly useful when the gene of interest plays a key role in the early development of many tissues, and a complete deletion of the gene from the germ line would cause death very early in embryogenesis. The same strategy can also be used to inappropriately express any specific gene in a tissue of interest; here, the triggered deletion joins a strong transcriptional promoter to the gene of interest. With this tool one can in principle determine the influence of any protein in any desired tissue of an intact animal.

Summary

The genomes of nearly all organisms contain mobile genetic elements that can move from one position in the genome to another by either transpositional or conservative site-specific recombination processes. In most cases, this movement is random and happens at a very low frequency. Mobile genetic elements include transposons, which move within a single cell (and its descendants), plus those viruses whose genomes can integrate into the genomes of their host cells.

There are three classes of transposons: the DNA-only transposons, the retroviral-like retrotransposons, and the nonretroviral retrotransposons. All but the last have close relatives among the viruses. Although viruses and transposable elements can be viewed as parasites, many of the new arrangements of DNA sequences that their site-specific recombination events produce have created the genetic variation crucial for the evolution of cells and organisms.

PROBLEMS

Which statements are true? Explain why or why not.

5–1 No two cells in your body have the identical nucleotide sequence.

5–2 In *E. coli*, where the replication fork travels at 500 nucleotide pairs per second, the DNA ahead of the fork must rotate at nearly 3000 revolutions per minute.

5–3 When bidirectional replication forks from adjacent origins meet, a leading strand always runs into a lagging strand.

5–4 DNA repair mechanisms all depend on the existence of two copies of the genetic information, one in each of the two homologous chromosomes.

Discuss the following problems.

5–5 To determine the reproducibility of mutation frequency measurements, you do the following experiment. You inoculate each of 10 cultures with a single *E. coli* bacterium, allow the cultures to grow until each contains 10^6 cells, and then measure the number of cells in each culture that carry a mutation in your gene of interest. You were so surprised by the initial results that you repeated the experiment to confirm them. Both sets of results display the same extreme variability, as shown in **Table Q5–1**. Assuming that the rate of mutation is constant, why do you suppose there is so much variation in the frequencies of mutant cells in different cultures?

Table Q5–1 Frequencies of mutant cells in multiple cultures (Problem 5–5).

	CULTURE (mutant cells/10^6 cells)									
EXPERIMENT	1	2	3	4	5	6	7	8	9	10
1	4	0	257	1	2	32	0	0	2	1
2	128	0	1	4	0	0	66	5	0	2

5–6 DNA repair enzymes preferentially repair mismatched bases on the newly synthesized DNA strand, using the old DNA strand as a template. If mismatches were repaired instead without regard for which strand served as template, would mismatch repair reduce replication errors? Would such an indiscriminate mismatch repair result in fewer mutations, more mutations, or the same number of mutations as there would have been without any repair at all? Explain your answers.

5–7 If DNA polymerase requires a perfectly paired primer in order to add the next nucleotide, how is it that any mismatched nucleotides "escape" the polymerase and become substrates for mismatch repair enzymes?

5–8 The laboratory you joined is studying the life cycle of an animal virus that uses a circular, double-stranded DNA as its genome. Your project is to define the location of the origin(s) of replication and to determine whether replication proceeds in one or both directions away from an origin (unidirectional or bidirectional replication). To accomplish your goal, you isolated replicating molecules, cleaved them with a restriction nuclease that cuts the viral genome at one site to produce a linear molecule from the circle, and examined the resulting molecules in the electron microscope. Some of the molecules you observed are illustrated schematically in **Figure Q5–1**. (Note that it is impossible to distinguish the orientation of one DNA molecule from another in the electron microscope.)

You must present your conclusions to the rest of the lab tomorrow. How will you answer the two questions your advisor posed for you? Is there a single, unique origin of replication or several origins? Is replication unidirectional or bidirectional?

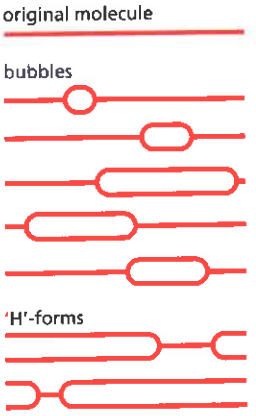


Figure Q5–1 Parental and replicating forms of an animal virus (Problem 5–8).

5–9 If you compare the frequency of the sixteen possible dinucleotide sequences in *E. coli* and human cells, there are no striking differences except for one dinucleotide, 5′-CG-3′. The frequency of CG dinucleotides in the human genome is significantly lower than in *E. coli* and significantly lower than expected by chance. Why do you suppose that CG dinucleotides are underrepresented in the human genome?

5–10 With age, somatic cells are thought to accumulate genomic "scars" as a result of the inaccurate repair of double-strand breaks by nonhomologous end-joining (NHEJ). Estimates based on the frequency of breaks in primary human fibroblasts suggest that by age 70 each human somatic cell may carry some 2000 NHEJ-induced mutations due to inaccurate repair. If these mutations were distributed randomly around the genome, how many genes would you expect to be affected? Would you expect cell function to be compromised? Why or why not? (Assume that 2% of the genome—1.5% coding and 0.5% regulatory—is crucial information.)

5–11 Draw the structure of the double Holliday junction that would result from strand invasion by both ends of the broken duplex into the intact homologous duplex shown in **Figure Q5–2**. Label the left end of each strand in the Holliday junction 5′ or 3′ so that the relationship to the parental and recombinant duplexes is clear. Indicate how DNA synthesis would be used to fill in any single-strand gaps in your double Holliday junction.

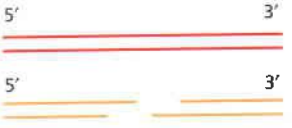


Figure Q5–2 A broken duplex with single-strand tails ready to invade an intact homologous duplex (Problem 5–11).

5–12 Why is it that recombination between similar, but nonidentical, repeated sequences poses a problem for human cells? How does the mismatch-repair system protect against such recombination events?

REFERENCES

General

Biological Response to DNA Damage (2000) *Cold Spring Harb Symp Quant Biol* 65.

Brown TA (2002) Genomes 2, 2nd ed. New York: Wiley-Liss.

Friedberg EC, Walker GC, Siede W et al (2005) DNA Repair and Mutagenesis. Washington, DC: ASM Press.

Hartwell L, Hood L, Goldberg ML et al (2006) Genetics: from Genes to Genomes. Boston: McGraw Hill.

Stent GS (1971) Molecular Genetics: An Introductory Narrative. San Francisco: WH Freeman.

Watson J, Baker T, Bell S et al (2004) Molecular Biology of the Gene, 5th ed. Plainview, New York: Cold Spring Harbor Laboratory Press.

The Maintenance of DNA Sequences

Cooper GM, Brudno M, Stone ES et al (2004) Characterization of evolutionary rates and constraints in three mammalian genomes. *Genome Res* 14:539–548.

Crow JF (2000) The origins, patterns and implications of human spontaneous mutation. *Nature Rev Genet* 1:40–47.

Hedges SB (2002) The origin and evolution of model organisms. *Nature Rev Genet* 3:838–849.

King MC, Wilson AC (1965) Evolution at two levels in humans and chimpanzees. *Science* 188:107–16.

DNA Replication Mechanisms

Alberts B (1998) The cell as a collection of protein machines: preparing the next generation of molecular biologists. *Cell* 92:291–294.

Dillingham MS (2006) Replicative helicases: a staircase with a twist. *Curr Biol* 16:R844–R847.

Indiani C & O'Donnell M (2006) The replication clamp-loading machine at work in the three domains of life. *Nature Rev Mol Cell Biol* 7:751–761.

Kornberg A (1960) Biological synthesis of DNA. *Science* 131:1503–1508.

Li JJ & Kelly TJ (1984) SV40 DNA replication *in vitro*. *Proc Natl Acad Sci USA* 81:6973.

Meselson M & Stahl FW (1958) The replication of DNA in *E. coli*. *Proc Natl Acad Sci USA* 44:671–682.

Modrich P & Lahue R (1996) Mismatch repair in replication fidelity, genetic recombination, and cancer biology. *Annu Rev Biochem* 65:101–133.

Mott ML & Berger JM (2007) DNA replication initiation: mechanisms and regulation in bacteria. *Nature Rev Microbiol* 5:343–354.

O'Donnell M (2006) Replisome architecture and dynamics in *E. coli*. *J Biol Chem* 281:10653–10656.

Okazaki R, Okazaki T, Sakabe K et al. (1968) Mechanism of DNA chain growth. I. Possible discontinuity and unusual secondary structure of newly synthesized chains. *Proc Natl Acad Sci USA* 59:598–605.

Raghuraman MK, Winzeler EA, Collingwood D et al (2001) Replication dynamics of the yeast genome. *Science* 294:115–121.

Rao PN & Johnson RT (1970) Mammalian cell fusion: studies on the regulation of DNA synthesis and mitosis. *Nature* 225:159.

Wang JC (2002) Cellular roles of DNA topoisomerases: a molecular perspective. *Nature Rev Mol Cell Biol* 3:430–440.

The Initiation and Completion of DNA Replication in Chromosomes

Chan SR & Blackburn EH (2004) Telomeres and telomerase. *Philos Trans R Soc Lond B Bio Sci* 359:109–121.

Costa S & Blow JJ (2007) The elusive determinants of replication origins. *EMBO Rep* 8:332–334.

Groth A, Rocha W & Almouzni G (2007) Chromatin challenges during DNA replication and repair. *Cell* 128:721–733.

Machida YJ, Hamlin JL & Dutta A (2005) Right place, right time, and only once: replication initiation in metazoans. *Cell* 123:13–24.

O'Donnell M & Kuriyan J (2005) Clamp loaders and replication initiation. *Curr Opin Struct Biol* 16:35–41.

Robinson NP & Bell SD (2005) Origins of DNA replication in the three domains of life. *FEBS J* 272:3757–3766.

Smogorzewska A & de Lange T (2004) Regulation of telomerase by telomeric proteins. *Annu Rev Biochem* 73:177–208.

DNA Repair

Barnes DE, & Lindahl T (2004) Repair and genetic consequences of endogenous DNA base damage in mammalian cells. *Annu Rev Genet* 38:445–476.

Harrison JC & Haber JE (2006) Surviving the breakup: the DNA damage checkpoint. *Annu Rev Genet* 40:209–235.

Heller RC & Marians KJ (2006) Replisome assembly and the direct restart of stalled replication. *Nature Rev Mol Cell Biol* 7:932–43.

Lieber M, Ma Y et al (2003) Mechanism and regulation of human non-homologous DNA end-joining. *Nature Rev Mol Cell Biol* 4:712–720.

Lindahl T (1993) Instability and decay of the primary structure of DNA. *Nature* 362:709–715.

Prakash S & Prakash L (2002) Translesion DNA synthesis in eukaryotes: a one- or two-polymerase affair. *Genes Dev* 16:1872–1883.

Sancar A, Lindsey-Boltz LA et al. (2004) Molecular mechanisms of mammalian DNA repair and the DNA damage checkpoints. *Annu Rev Biochem* 73:39–85.

Svejstrup JQ (2002) Mechanisms of transcription-coupled DNA repair. *Nature Rev Mol Cell Biol* 3:21–29.

Wyman C & Kanaar R (2006) DNA double-strand break repair: all's well that ends well. *Annu Rev Genet* 40:363–383.

Homologous Recombination

Adams MD, McVey M & Sekelsky JJ (2003) *Drosophila* BLM in double-strand break repair by synthesis-dependent strand annealing. *Science* 299:265–267.

Cox MM (2001) Historical overview: searching for replication help in all of the rec places. *Proc Natl Acad Sci USA* 98:8173–8180.

Holliday R (1990) The history of the DNA heteroduplex. *BioEssays* 12:133–142.

Lisby M, Bartow JH et al (2004) Choreography of the DNA damage response: spatiotemporal relationships among checkpoint and repair proteins. *Cell* 118:699–713.

McEachern MJ & Haber JE (2006) Break-induced replication and recombinational telomere elongation in yeast. *Annu Rev Biochem* 75:111–135.

Michel B, Gromponee G et al (2004) Multiple pathways process stalled replication forks. *Proc Natl Acad Sci USA* 101:12783–12788.

Szostak JW, Orr-Weaver TK, Rothstein RJ et al (1983) The double-strand break repair model for recombination. *Cell* 33:25–35.

West SC (2003) Molecular views of recombination proteins and their control. *Nature Rev Mol Cell Biol* 4:435–445.

Transposition and Site-Specific Recombination

Campbell AM (1993) Thirty years ago in genetics: prophage insertion into bacterial chromosomes. *Genetics* 133:433–438.

Comfort NC (2001) From controlling elements to transposons: Barbara McClintock and the Nobel Prize. *Trends Biochem Sci* 26:454–457.

Craig NL (1996) Transposition, in *Escherichia coli* and *Salmonella*, pp 2339–2362. Washington, DC: ASM Press.

Gottesman M (1999) Bacteriophage lambda: the untold story. *J Mol Biol* 293:177–180.

Grindley ND, Whiteson KL & Rice PA (2006) Mechanisms of site-specific recombination. *Annu Rev Biochem* 75:567–605.

Varmus H (1988) Retroviruses. *Science* 240:1427–1435.

Zickler D & Kleckner N (1999) Meiotic chromosomes: integrating structure and function. *Annu Rev Genet* 33:603–754.

How Cells Read the Genome: From DNA to Protein

6

In This Chapter

FROM DNA TO RNA 331

FROM RNA TO PROTEIN 366

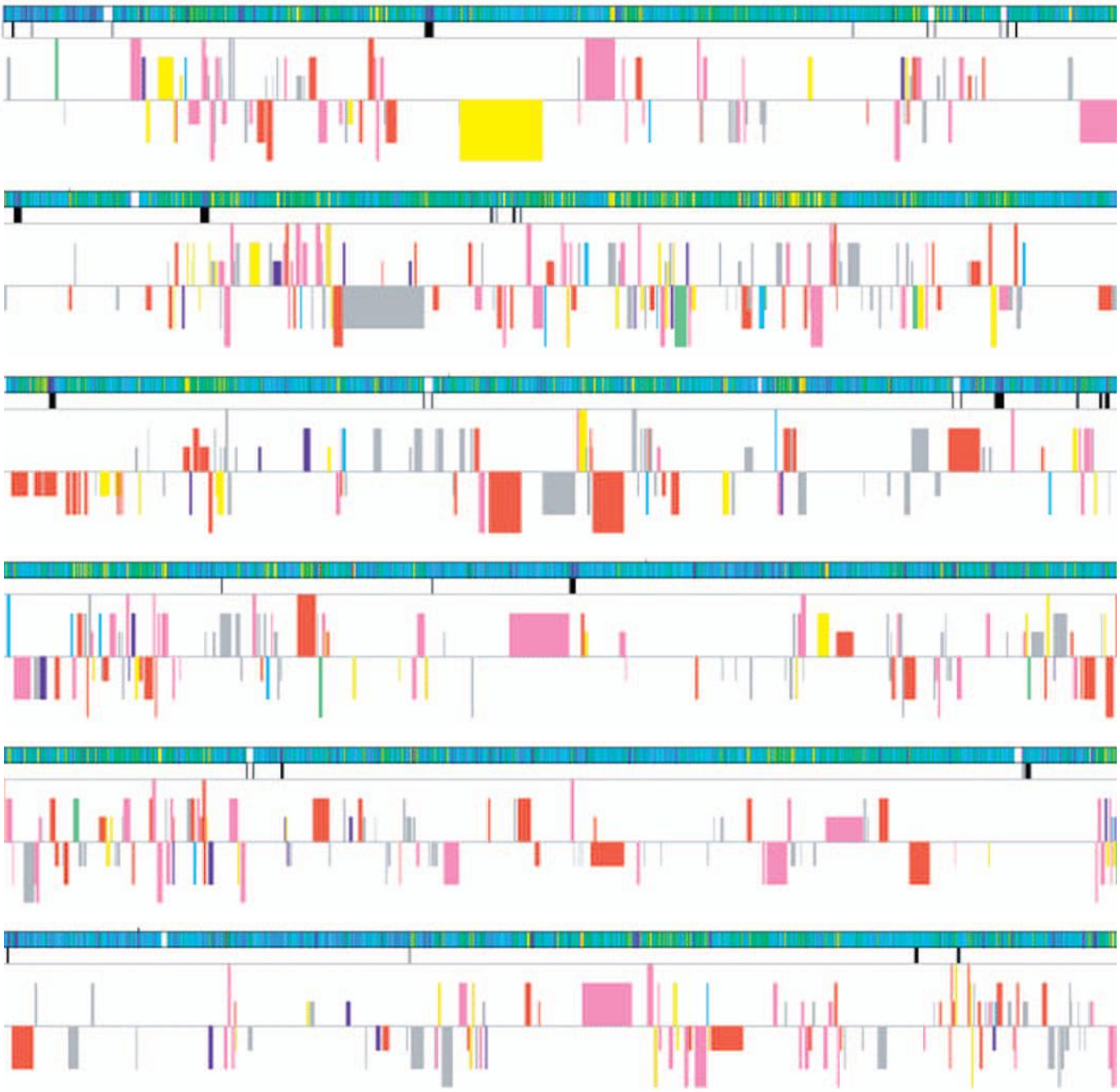
THE RNA WORLD AND THE ORIGINS OF LIFE 400

Only when the structure of DNA was discovered in the early 1950s did it become clear how the hereditary information in cells is encoded in DNA's sequence of nucleotides. The progress since then has been astounding. Within fifty years we knew the complete genome sequences for many organisms, including humans. We therefore know the maximum amount of information that is required to produce a complex organism like ourselves. The limits on the hereditary information needed for life constrain the biochemical and structural features of cells and make it clear that biology is not infinitely complex.

In this chapter, we explain how cells decode and use the information in their genomes. Much has been learned about how the genetic instructions written in an alphabet of just four "letters"—the four different nucleotides in DNA—direct the formation of a bacterium, a fruit fly, or a human. Nevertheless, we still have a great deal to discover about how the information stored in an organism's genome produces even the simplest unicellular bacterium with 500 genes, let alone how it directs the development of a human with approximately 25,000 genes. An enormous amount of ignorance remains; many fascinating challenges therefore await the next generation of cell biologists.

The problems that cells face in decoding genomes can be appreciated by considering a small portion of the genome of the fruit fly *Drosophila melanogaster* (**Figure 6–1**). Much of the DNA-encoded information present in this and other genomes specifies the linear order—the sequence—of amino acids for every protein the organism makes. As described in Chapter 3, the amino acid sequence in turn dictates how each protein folds to give a molecule with a distinctive shape and chemistry. When a cell makes a particular protein, it must decode accurately the corresponding region of the genome. Additional information encoded in the DNA of the genome specifies exactly when in the life of an organism and in which cell types each gene is to be expressed into protein. Since proteins are the main constituents of cells, the decoding of the genome determines not only the size, shape, biochemical properties, and behavior of cells, but also the distinctive features of each species on Earth.

One might have predicted that the information present in genomes would be arranged in an orderly fashion, resembling a dictionary or a telephone directory. Although the genomes of some bacteria seem fairly well organized, the genomes of most multicellular organisms, such as our *Drosophila* example, are surprisingly disorderly. Small bits of coding DNA (that is, DNA that codes for protein) are interspersed with large blocks of seemingly meaningless DNA. Some sections of the genome contain many genes and others lack genes altogether. Proteins that work closely with one another in the cell often have their genes located on different chromosomes, and adjacent genes typically encode proteins that have little to do with each other in the cell. Decoding genomes is therefore no simple matter. Even with the aid of powerful computers, it is still difficult for researchers to locate definitively the beginning and end of genes in the DNA sequences of complex genomes, much less to predict when each gene is expressed in the life of the organism. Although the DNA sequence of the human genome is known, it will probably take at least a decade to identify every gene and determine the precise amino acid sequence of the protein it produces. Yet the cells in our body do this thousands of times a second.



KEY:

13 or more
1 to 12
none

length of bar indicates number of corresponding cDNAs identified in databases

100,000 nucleotide pairs

%GC content 25 65

transposable elements

known and predicted genes identified on top strand of DNA

known and predicted genes identified on bottom strand of DNA

color code for sequence similarity of genes identified

MWY
WY
MW
W
MY
Y
M
no similarity to MWY

M = mammalian
W = *C. elegans*
Y = *S. cerevisiae*

Figure 6–1 ***(opposite page)*** **Schematic depiction of a portion of chromosome 2 from the genome of the fruit fly *Drosophila melanogaster*.** This figure represents approximately 3% of the total *Drosophila* genome, arranged as six contiguous segments. As summarized in the key, the symbolic representations are: *black vertical lines* of various thicknesses: locations of transposable elements, with thicker bars indicating clusters of elements; *colored boxes*: genes (both known and predicted) coded on one strand of DNA (boxes *above* the midline) and genes coded on the other strand (boxes *below* the midline). The length of each gene box includes both its exons (protein-coding DNA) and its introns (noncoding DNA) (see Figure 4–15); its height is proportional to the number of known cDNAs that match the gene. (As described in Chapter 8, cDNAs are DNA copies of mRNA molecules, and large collections of the nucleotide sequences of cDNAs have been deposited in a variety of databases, the more matches, the higher the confidence that the predicted gene is transcribed into RNA and is thus a genuine gene.) The color of each gene box indicates whether a closely related gene is known to occur in other organisms. For example, MWY means the gene has close relatives in mammals, in the nematode worm *Caenorhabditis elegans*, and in the yeast *Saccharomyces cerevisiae*. MW indicates the gene has close relatives in mammals and the worm but not in yeast. The *rainbow-colored bar* indicates percent G–C base pairs; across many different genomes, this percentage shows a striking regional variation, whose origin and significance are uncertain. (From M.D. Adams et al., *Science* 287:2185–2195, 2000. With permission from AAAS.)

The DNA in genomes does not direct protein synthesis itself, but instead uses RNA as an intermediary. When the cell needs a particular protein, the nucleotide sequence of the appropriate portion of the immensely long DNA molecule in a chromosome is first copied into RNA (a process called *transcription)*. It is these RNA copies of segments of the DNA that are used directly as templates to direct the synthesis of the protein (a process called *translation)*. The flow of genetic information in cells is therefore from DNA to RNA to protein (**Figure 6–2**). All cells, from bacteria to humans, express their genetic information in this way—a principle so fundamental that it is termed the *central dogma* of molecular biology.

Despite the universality of the central dogma, there are important variations in the way in which information flows from DNA to protein. Principal among these is that RNA transcripts in eucaryotic cells are subject to a series of processing steps in the nucleus, including *RNA splicing*, before they are permitted to exit from the nucleus and be translated into protein. These processing steps can critically change the "meaning" of an RNA molecule and are therefore crucial for understanding how eucaryotic cells read their genomes. Finally, although we focus on the production of the proteins encoded by the genome in this chapter, we see that for many genes RNA is the final product. Like proteins, many of these RNAs fold into precise three-dimensional structures that have structural, catalytic, and regulatory roles in the cell.

We begin this chapter with the first step in decoding a genome: the process of transcription by which an RNA molecule is produced from the DNA of a gene. We then follow the fate of this RNA molecule through the cell, finishing when a correctly folded protein molecule has been formed. At the end of the chapter, we consider how the present quite complex scheme of information storage, transcription, and translation might have arisen from simpler systems in the earliest stages of cell evolution.

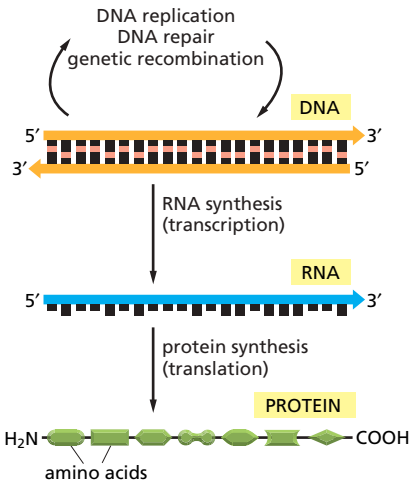


Figure 6–2 The pathway from DNA to protein. The flow of genetic information from DNA to RNA (transcription) and from RNA to protein (translation) occurs in all living cells.

FROM DNA TO RNA

Transcription and translation are the means by which cells read out, or express, the genetic instructions in their genes. Because many identical RNA copies can be made from the same gene, and each RNA molecule can direct the synthesis of many identical protein molecules, cells can synthesize a large amount of protein rapidly when necessary. But each gene can also be transcribed and translated with a different efficiency, allowing the cell to make vast quantities of some proteins and tiny quantities of others (**Figure 6–3**). Moreover, as we see in the next chapter, a cell can change (or regulate) the expression of each of its genes according to the needs of the moment—most commonly by controlling the production of its RNA.

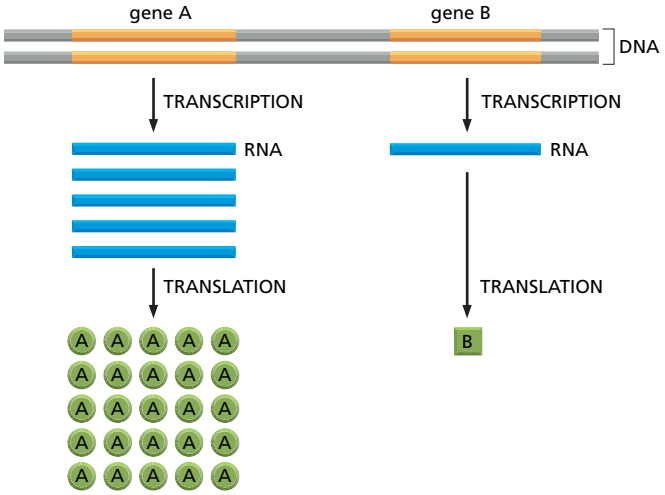


Figure 6–3 Genes can be expressed with different efficiencies. In this example, gene A is transcribed and translated much more efficiently than gene B. This allows the amount of protein A in the cell to be much greater than that of protein B.

Portions of DNA Sequence Are Transcribed into RNA

The first step a cell takes in reading out a needed part of its genetic instructions is to copy a particular portion of its DNA nucleotide sequence—a gene—into an RNA nucleotide sequence. The information in RNA, although copied into another chemical form, is still written in essentially the same language as it is in DNA—the language of a nucleotide sequence. Hence the name **transcription**.

Like DNA, RNA is a linear polymer made of four different types of nucleotide subunits linked together by phosphodiester bonds (**Figure 6–4**). It differs from DNA chemically in two respects: (1) the nucleotides in RNA are *ribonucleotides*—that is, they contain the sugar ribose (hence the name *ribo*nucleic acid) rather than deoxyribose; (2) although, like DNA, RNA contains the bases adenine (A), guanine (G), and cytosine (C), it contains the base uracil (U) instead of the thymine (T) in DNA. Since U, like T, can base-pair by hydrogen-bonding with A (**Figure 6–5**), the complementary base-pairing properties described for DNA in Chapters 4 and 5 apply also to RNA (in RNA, G pairs with C, and A pairs with U). We also find other types of base pairs in RNA: for example, G occasionally pairs with U.

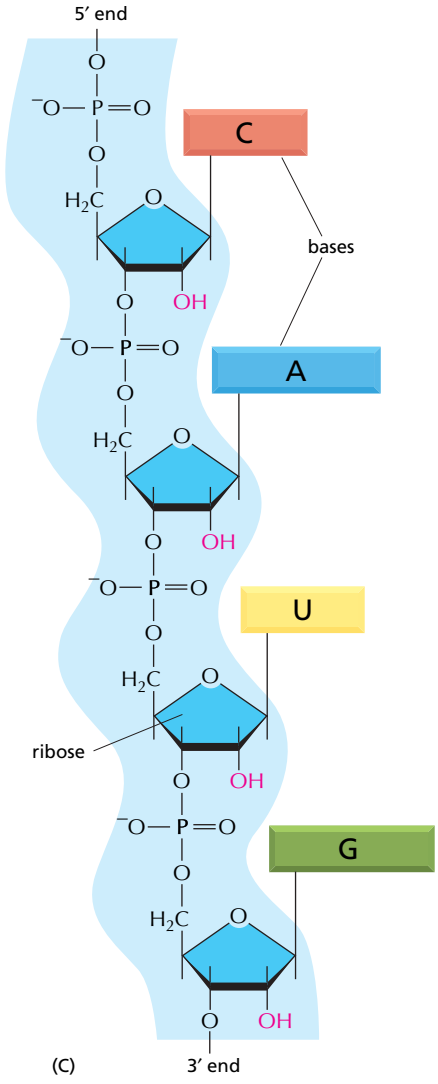
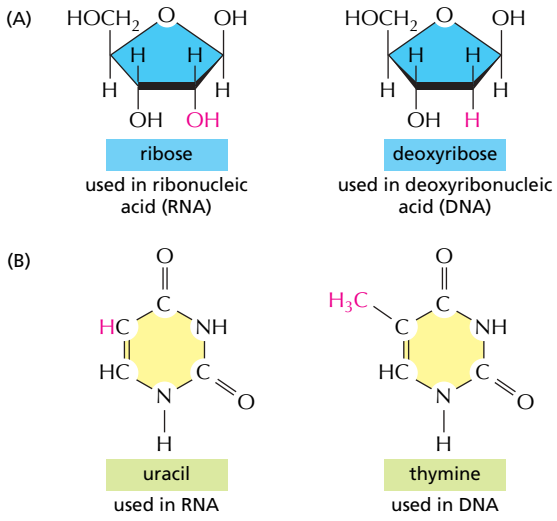


Figure 6–4 The chemical structure of RNA. (A) RNA contains the sugar ribose, which differs from deoxyribose, the sugar used in DNA, by the presence of an additional –OH group. (B) RNA contains the base uracil, which differs from thymine, the equivalent base in DNA, by the absence of a $-CH_3$ group. (C) A short length of RNA. The phosphodiester chemical linkage between nucleotides in RNA is the same as that in DNA.



Although these chemical differences are slight, DNA and RNA differ quite dramatically in overall structure. Whereas DNA always occurs in cells as a double-stranded helix, RNA is single-stranded. An RNA chain can therefore fold up into a particular shape, just as a polypeptide chain folds up to form the final shape of a protein (**Figure 6–6**). As we see later in this chapter, the ability to fold into complex three-dimensional shapes allows some RNA molecules to have precise structural and catalytic functions.

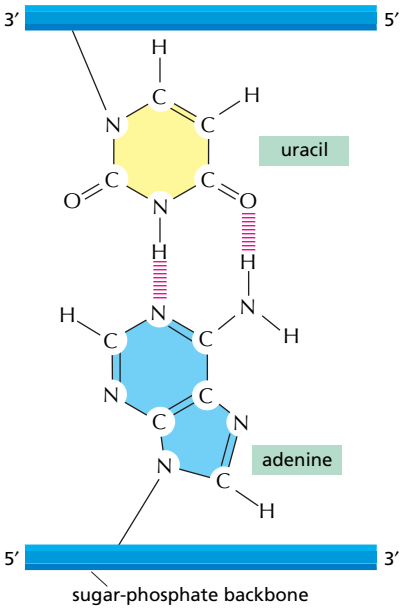


Figure 6–5 Uracil forms base pairs with adenine. The absence of a methyl group in U has no effect on base-pairing; thus, U–A base pairs closely resemble T–A base pairs (see Figure 4–4).

Transcription Produces RNA Complementary to One Strand of DNA

The RNA in a cell is made by DNA transcription, a process that has certain similarities to the process of DNA replication discussed in Chapter 5. Transcription begins with the opening and unwinding of a small portion of the DNA double helix to expose the bases on each DNA strand. One of the two strands of the DNA double helix then acts as a template for the synthesis of an RNA molecule. As in DNA replication, the nucleotide sequence of the RNA chain is determined by the complementary base-pairing between incoming nucleotides and the DNA template. When a good match is made, the incoming ribonucleotide is covalently linked to the growing RNA chain in an enzymatically catalyzed reaction. The RNA chain produced by transcription—the *transcript*—is therefore elongated one nucleotide at a time, and it has a nucleotide sequence that is exactly complementary to the strand of DNA used as the template (**Figure 6–7**).

Transcription, however, differs from DNA replication in several crucial ways. Unlike a newly formed DNA strand, the RNA strand does not remain hydrogen-bonded to the DNA template strand. Instead, just behind the region where the ribonucleotides are being added, the RNA chain is displaced and the DNA helix re-forms. Thus, the RNA molecules produced by transcription are released from the DNA template as single strands. In addition, because they are copied from only a limited region of the DNA, RNA molecules are much shorter than DNA molecules. A DNA molecule in a human chromosome can be up to 250 million nucleotide-pairs long; in contrast, most RNAs are no more than a few thousand nucleotides long, and many are considerably shorter.

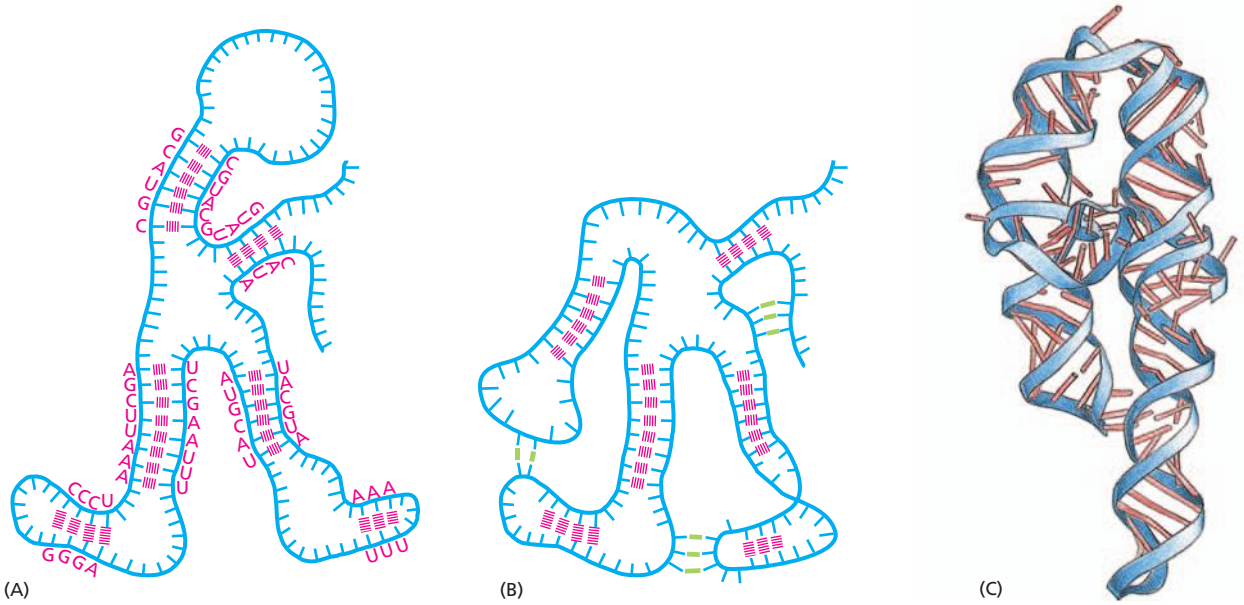


Figure 6–6 RNA can fold into specific structures. RNA is largely single-stranded, but it often contains short stretches of nucleotides that can form conventional base pairs with complementary sequences found elsewhere on the same molecule. These interactions, along with additional "nonconventional" base-pair interactions, allow an RNA molecule to fold into a three-dimensional structure that is determined by its sequence of nucleotides. <AATC> (A) Diagram of a folded RNA structure showing only conventional base-pair interactions. (B) Structure with both conventional *(red)* and nonconventional *(green)* base-pair interactions. (C) Structure of an actual RNA, a portion of a group I intron (see Figure 6–36). Each conventional base-pair interaction is indicated by a "rung" in the double helix. Bases in other configurations are indicated by broken rungs.

The enzymes that perform transcription are called **RNA polymerases**. Like the DNA polymerase that catalyzes DNA replication (discussed in Chapter 5), RNA polymerases catalyze the formation of the phosphodiester bonds that link the nucleotides together to form a linear chain. The RNA polymerase moves stepwise along the DNA, unwinding the DNA helix just ahead of the active site for polymerization to expose a new region of the template strand for complementary base-pairing. In this way, the growing RNA chain is extended by one nucleotide at a time in the 5′-to-3′ direction (**Figure 6–8**). The substrates are nucleoside triphosphates (ATP, CTP, UTP, and GTP); as in DNA replication, the hydrolysis of high-energy bonds provides the energy needed to drive the reaction forward (see Figure 5–4).

The almost immediate release of the RNA strand from the DNA as it is synthesized means that many RNA copies can be made from the same gene in a relatively short time, with the synthesis of additional RNA molecules being started before the first RNA is completed (**Figure 6–9**). When RNA polymerase molecules follow hard on each other's heels in this way, each moving at about 20 nucleotides per second (the speed in eucaryotes), over a thousand transcripts can be synthesized in an hour from a single gene.

Although RNA polymerase catalyzes essentially the same chemical reaction as DNA polymerase, there are some important differences between the activities of the two enzymes. First, and most obviously, RNA polymerase catalyzes the linkage of ribonucleotides, not deoxyribonucleotides. Second, unlike the DNA polymerases involved in DNA replication, RNA polymerases can start an RNA chain without a primer. This difference may exist because transcription need not be as accurate as DNA replication (see Table 5–1, p. 271). Unlike DNA, RNA does not permanently store genetic information in cells. RNA polymerases make about one mistake for every 10^4 nucleotides copied into RNA (compared with an error rate for direct copying by DNA polymerase of about one in 10^7 nucleotides), and the consequences of an error in RNA transcription are much less significant than that in DNA replication.

Although RNA polymerases are not nearly as accurate as the DNA polymerases that replicate DNA, they nonetheless have a modest proofreading mechanism. If an incorrect ribonucleotide is added to the growing RNA chain, the polymerase can back up, and the active site of the enzyme can perform an excision reaction that resembles the reverse of the polymerization reaction,

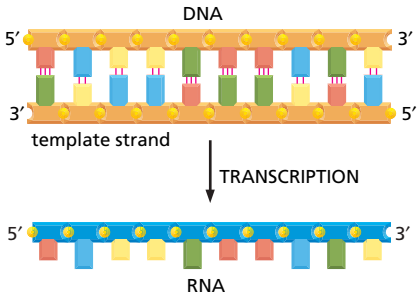


Figure 6–7 DNA transcription produces a single-stranded RNA molecule that is complementary to one strand of DNA.

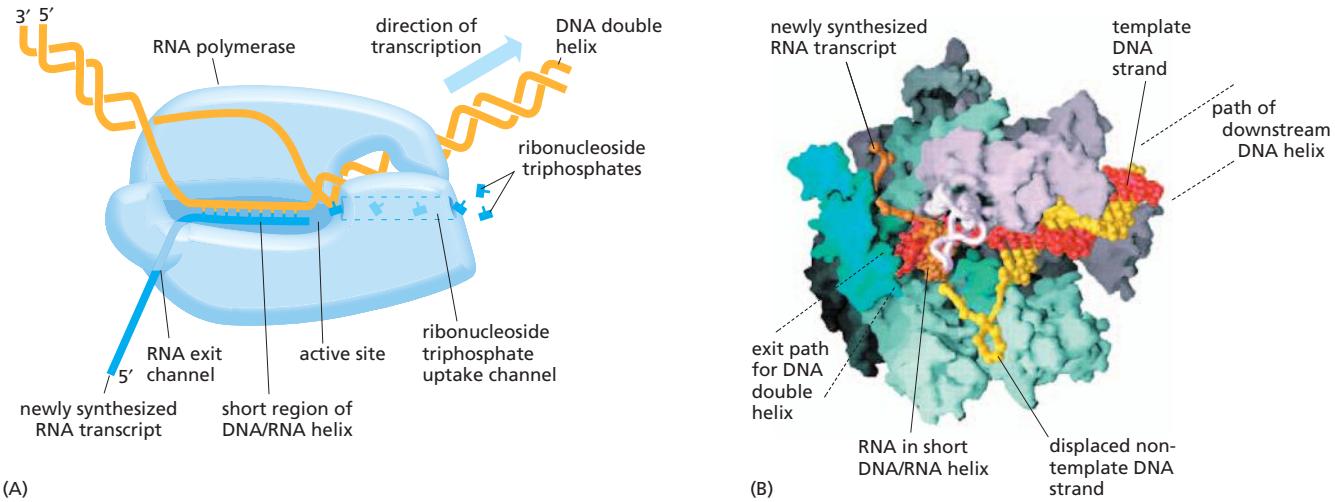


Figure 6–8 DNA is transcribed by the enzyme RNA polymerase. (A) The RNA polymerase *(pale blue)* moves stepwise along the DNA, unwinding the DNA helix at its active site. As it progresses, the polymerase adds nucleotides (represented as *small "T" shapes)* one by one to the RNA chain at the polymerization site, using an exposed DNA strand as a template. The RNA transcript is thus a complementary copy of one of the two DNA strands. A short region of DNA/RNA helix (approximately nine nucleotide pairs in length) is therefore formed only transiently, and a "window" of DNA/RNA helix therefore moves along the DNA with the polymerase. The incoming nucleotides are in the form of ribonucleoside triphosphates (ATP, UTP, CTP, and GTP), and the energy stored in their phosphate–phosphate bonds provides the driving force for the polymerization reaction (see Figure 5–4). (B) The structure of a bacterial RNA polymerase, as determined by x-ray crystallography. Four different subunits, indicated by different colors, comprise this RNA polymerase. The DNA strand used as a template is *red*, and the nontemplate strand is *yellow*. (A, adapted from a figure courtesy of Robert Landick; B, courtesy of Seth Darst.)

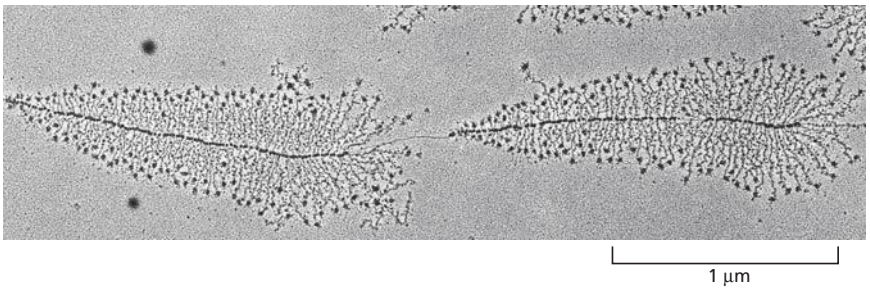


Figure 6–9 Transcription of two genes as observed under the electron microscope. The micrograph shows many molecules of RNA polymerase simultaneously transcribing each of two adjacent genes. Molecules of RNA polymerase are visible as a series of dots along the DNA with the newly synthesized transcripts (fine threads) attached to them. The RNA molecules (ribosomal RNAs) shown in this example are not translated into protein but are instead used directly as components of ribosomes, the machines on which translation takes place. The particles at the 5′ end (the free end) of each rRNA transcript are believed to reflect the beginnings of ribosome assembly. From the lengths of the newly synthesized transcripts, it can be deduced that the RNA polymerase molecules are transcribing from left to right. (Courtesy of Ulrich Scheer.)

except that water instead of pyrophosphate is used and a nucleoside monophosphate is released.

Given that DNA and RNA polymerases both carry out template-dependent nucleotide polymerization, it might be expected that the two types of enzymes would be structurally related. However, x-ray crystallographic studies of both types of enzymes reveal that, other than containing a critical Mg^{2+} ion at the catalytic site, they are virtually unrelated to each other; indeed template-dependent nucleotide polymerizing enzymes seem to have arisen independently twice during the early evolution of cells. One lineage led to the modern DNA polymerases and reverse transcriptases discussed in Chapter 5, as well as to a few single-subunit RNA polymerases from viruses. The other lineage formed all of the modern cellular RNA polymerases (**Figure 6–10**), which we discuss in this chapter.

Cells Produce Several Types of RNA

The majority of genes carried in a cell's DNA specify the amino acid sequence of proteins; the RNA molecules that are copied from these genes (which ultimately direct the synthesis of proteins) are called **messenger RNA** (**mRNA**) molecules. The final product of a minority of genes, however, is the RNA itself. Careful analysis of the complete DNA sequence of the genome of the yeast *S. cerevisiae* has uncovered well over 750 genes (somewhat more than 10% of the total number of yeast genes) that produce RNA as their final product. These RNAs, like proteins, serve as enzymatic and structural components for a wide variety of processes in

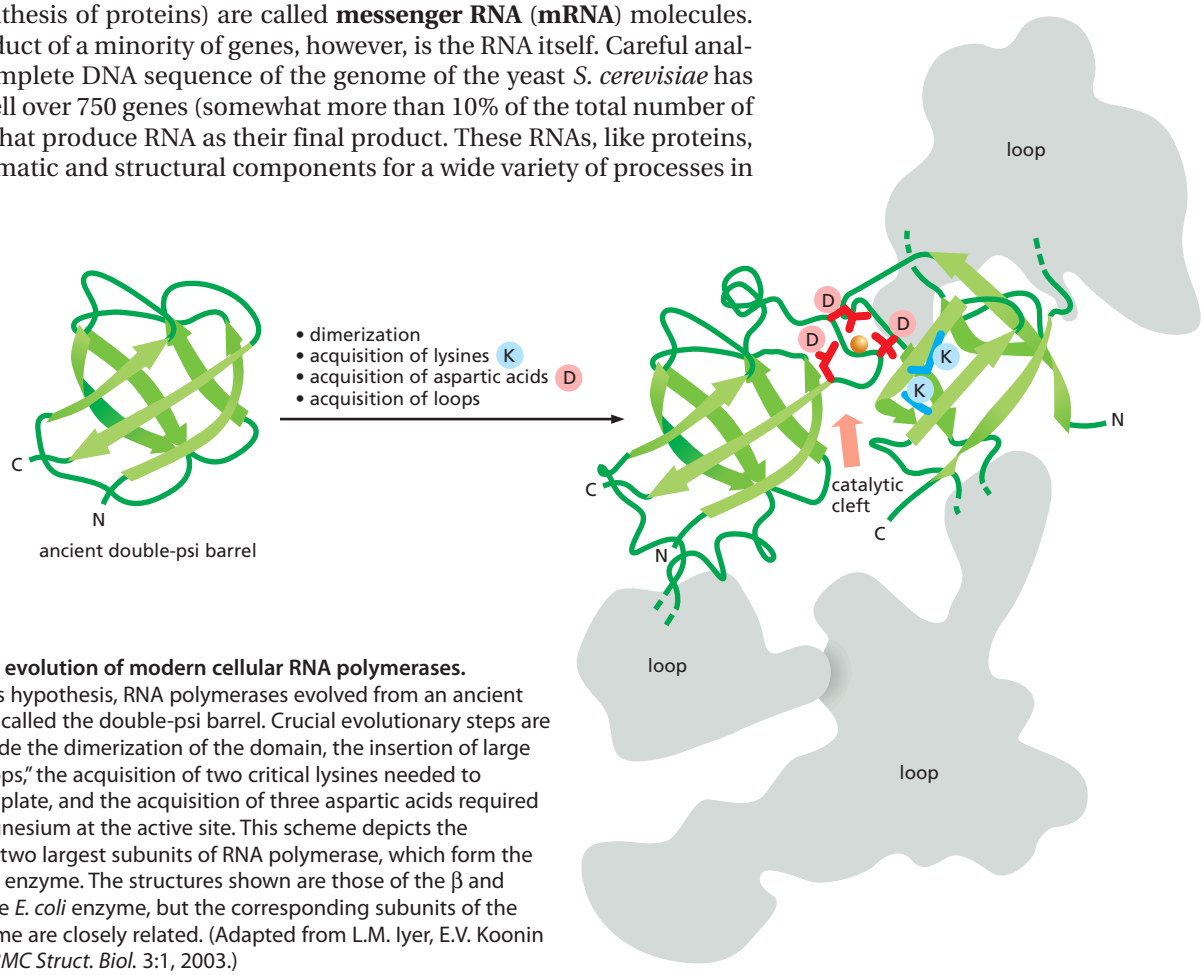


Figure 6–10 The evolution of modern cellular RNA polymerases. According to this hypothesis, RNA polymerases evolved from an ancient protein domain, called the double-psi barrel. Crucial evolutionary steps are thought to include the dimerization of the domain, the insertion of large polypeptide "loops," the acquisition of two critical lysines needed to position the template, and the acquisition of three aspartic acids required to chelate a magnesium at the active site. This scheme depicts the evolution of the two largest subunits of RNA polymerase, which form the active site of the enzyme. The structures shown are those of the β and β′ subunits of the *E. coli* enzyme, but the corresponding subunits of the eucaryotic enzyme are closely related. (Adapted from L.M. Iyer, E.V. Koonin and L. Aravind, *BMC Struct. Biol.* 3:1, 2003.)

the cell. In Chapter 5 we encountered one of those RNAs, the template carried by the enzyme telomerase. Although many of these noncoding RNAs are still mysterious, we shall see in this chapter that *small nuclear RNA (snRNA)* molecules direct the splicing of pre-mRNA to form mRNA, that *ribosomal RNA (rRNA)* molecules form the core of ribosomes, and that *transfer RNA (tRNA)* molecules form the adaptors that select amino acids and hold them in place on a ribosome for incorporation into protein. Finally, we shall see in Chapter 7 that *microRNA (miRNA)* molecules and *small interfering RNA (siRNA)* molecules serve as key regulators of eucaryotic gene expression (**Table 6–1**).

Each transcribed segment of DNA is called a *transcription unit*. In eucaryotes, a transcription unit typically carries the information of just one gene, and therefore codes for either a single RNA molecule or a single protein (or group of related proteins if the initial RNA transcript is spliced in more than one way to produce different mRNAs). In bacteria, a set of adjacent genes is often transcribed as a unit; the resulting mRNA molecule therefore carries the information for several distinct proteins.

Overall, RNA makes up a few percent of a cell's dry weight. Most of the RNA in cells is rRNA; mRNA comprises only 3–5% of the total RNA in a typical mammalian cell. The mRNA population is made up of tens of thousands of different species, and there are on average only 10–15 molecules of each species of mRNA present in each cell.

Signals Encoded in DNA Tell RNA Polymerase Where to Start and Stop

To transcribe a gene accurately, RNA polymerase must recognize where on the genome to start and where to finish. The way in which RNA polymerases perform these tasks differs somewhat between bacteria and eucaryotes. Because the processes in bacteria are simpler, we discuss them first.

The initiation of transcription is an especially important step in gene expression because it is the main point at which the cell regulates which proteins are to be produced and at what rate. The bacterial RNA polymerase core enzyme is a multisubunit complex that sythesizes RNA using a DNA template as a guide. A detachable subunit called *sigma (σ) factor* associates with the core enzyme and assists it in reading the signals in the DNA that tell it where to begin transcribing (**Figure 6–11**). Together, σ factor and core enzyme are known as the **RNA polymerase holoenzyme**; this complex adheres only weakly to bacterial DNA when

Table 6–1 Principal Types of RNAs Produced in Cells

TYPE OF RNA	FUNCTION
mRNAs	messenger RNAs, code for proteins
rRNAs	ribosomal RNAs, form the basic structure of the ribosome and catalyze protein synthesis
tRNAs	transfer RNAs, central to protein synthesis as adaptors between mRNA and amino acids
snRNAs	small nuclear RNAs, function in a variety of nuclear processes, including the splicing of pre-mRNA
snoRNAs	small nucleolar RNAs, used to process and chemically modify rRNAs
scaRNAs	small cajal RNAs, used to modify snoRNAs and snRNAs
miRNAs	microRNAs, regulate gene expression typically by blocking translation of selective mRNAs
siRNAs	small interfering RNAs, turn off gene expression by directing degradation of selective mRNAs and the establishment of compact chromatin structures
Other noncoding RNAs	function in diverse cell processes, including telomere synthesis, X-chromosome inactivation, and the transport of proteins into the ER

the two collide, and a holoenzyme typically slides rapidly along the long DNA molecule until it dissociates again. However, when the polymerase holoenzyme slides into a region on the DNA double helix called a **promoter**, a special sequence of nucleotides indicating the starting point for RNA synthesis, the polymerase binds tightly to this DNA. The polymerase holoenzyme, through its σ factor, recognizes the promoter DNA sequence by making specific contacts with the portions of the bases that are exposed on the outside of the helix (step 1 in Figure 6–11).

After the RNA polymerase holoenzyme binds tightly to the promoter DNA in this way, it opens up the double helix to expose a short stretch of nucleotides on each strand (step 2 in Figure 6–11). Unlike a DNA helicase reaction (see Figure 5–14), this limited opening of the helix does not require the energy of ATP hydrolysis. Instead, the polymerase and DNA both undergo reversible structural changes that result in a state more energetically favorable than that of the initial binding. With the DNA unwound, one of the two exposed DNA strands acts as a template for complementary base-pairing with incoming ribonucleotides, two of which are joined together by the polymerase to begin an RNA chain (step 3 in Figure 6–11). After the first ten or so nucleotides of RNA have been synthesized (a relatively inefficient process during which polymerase synthesizes and discards short RNA oligomers), the core enzyme breaks its interactions with the promoter DNA, weakens its interactions with σ factor, and begins to move down the DNA, synthesizing RNA (steps 4 and 5 in Figure 6–11). Chain elongation continues (at a speed of approximately 50 nucleotides/sec for bacterial RNA polymerases) until the enzyme encounters a second signal in the DNA, the **terminator** (described below), where the polymerase halts and releases both the newly made RNA chain and the DNA template (step 7 in Figure 6–11). After the polymerase core enzyme has been released at a terminator, it reassociates with a free σ factor to form a holoenzyme that can begin the process of transcription again.

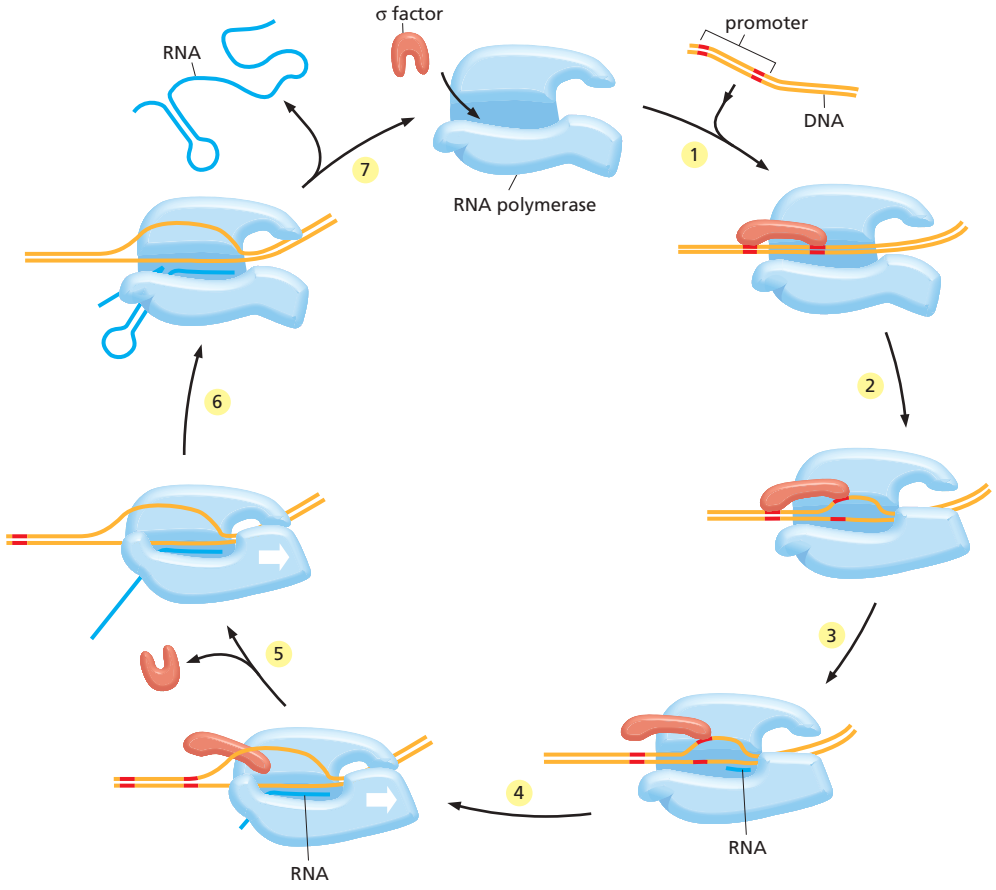
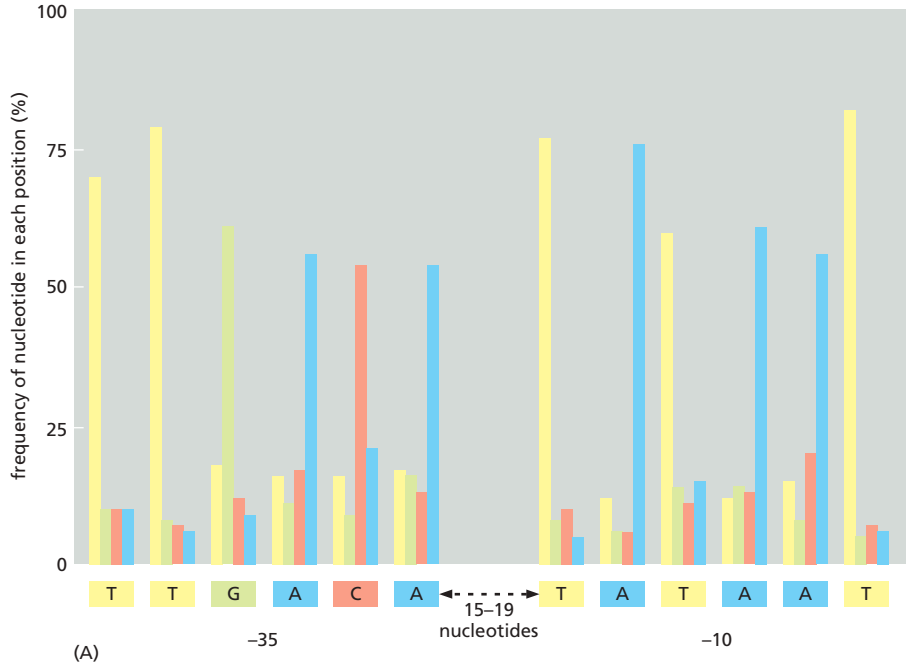


Figure 6–11 The transcription cycle of bacterial RNA polymerase. In step 1, the RNA polymerase holoenzyme (polymerase core enzyme plus σ factor) assembles and then locates a promoter (see Figure 6–12). The polymerase unwinds the DNA at the position at which transcription is to begin (step 2) and begins transcribing (step 3). This initial RNA synthesis (sometimes called "abortive initiation") is relatively inefficient. However, once RNA polymerase has managed to synthesize about 10 nucleotides of RNA, it breaks its interactions with the promoter DNA and weakens, and eventually breaks, its interaction with σ. The polymerase now shifts to the elongation mode of RNA synthesis (step 4), moving rightward along the DNA in this diagram. During the elongation mode (step 5), transcription is highly processive, with the polymerase leaving the DNA template and releasing the newly transcribed RNA only when it encounters a termination signal (steps 6 and 7). Termination signals are typically encoded in DNA, and many function by forming an RNA structure that destabilizes the polymerase's hold on the RNA (step 7). In bacteria, all RNA molecules are synthesized by a single type of RNA polymerase and the cycle depicted in the figure therefore applies to the production of mRNAs as well as structural and catalytic RNAs. (Adapted from a figure courtesy of Robert Landick.)



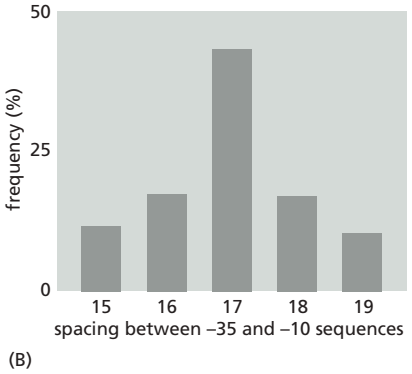


Figure 6–12 Consensus sequence for the major class of *E. coli* promoters. (A) The promoters are characterized by two hexameric DNA sequences, the –35 sequence and the –10 sequence named for their approximate location relative to the start point of transcription (designated +1). For convenience, the nucleotide sequence of a single strand of DNA is shown; in reality the RNA polymerase recognizes the promoter as double-stranded DNA. On the basis of a comparison of 300 promoters, the frequencies of the four nucleotides at each position in the –35 and –10 hexamers are given. The consensus sequence, shown *below* the graph, reflects the most common nucleotide found at each position in the collection of promoters. The sequence of nucleotides between the –35 and –10 hexamers shows no significant similarities among promoters. (B) The distribution of spacing between the –35 and –10 hexamers found in *E. coli* promoters.

The information displayed in these two graphs applies to *E. coli* promoters that are recognized by RNA polymerase and the major σ factor (designated σ^{70}). As we shall see in the next chapter, bacteria also contain minor σ factors, each of which recognizes a different promoter sequence. Some particularly strong promoters recognized by RNA polymerase and σ^{70} have an additional sequence, located upstream (to the *left,* in the figure) of the –35 hexamer, which is recognized by another subunit of RNA polymerase.

The process of transcription initiation is complex and requires that the RNA polymerase holoenzyme and the DNA undergo a series of conformational changes. We can view these changes as opening up and positioning the DNA in the active site followed by a successive tightening of the enzyme around the DNA and RNA to ensure that it does not dissociate before it has finished transcribing a gene. If an RNA polymerase does dissociate prematurely, it cannot resume synthesis but must start over again at the promoter.

How do the termination signals in the DNA stop the elongating polymerase? For most bacterial genes a termination signal consists of a string of A–T nucleotide pairs preceded by a two-fold symmetric DNA sequence, which, when transcribed into RNA, folds into a "hairpin" structure through Watson–Crick base-pairing (see Figure 6–11). As the polymerase transcribes across a terminator, the formation of the hairpin may help to "pull" the RNA transcript from the active site. The DNA–RNA hybrid in the active site, which is held together at terminators predominantly by U–A base pairs (which are less stable than G–C base pairs because they form two rather than three hydrogen bonds per base pair), is not strong enough to hold the RNA in place, and it dissociates causing the release of the polymerase from the DNA (step 7 in Figure 6–11). Thus, in some respects, transcription termination seems to involve a reversal of the structural transitions that happen during initiation. The process of termination also is an example of a common theme in this chapter: the folding of RNA into specific structures affects many steps in decoding the genome.

Transcription Start and Stop Signals Are Heterogeneous in Nucleotide Sequence

As we have just seen, the processes of transcription initiation and termination involve a complicated series of structural transitions in protein, DNA, and RNA molecules. The signals encoded in DNA that specify these transitions are often difficult for researchers to recognize. Indeed, a comparison of many different bacterial promoters reveals a surprising degree of variation. Nevertheless, they all contain related sequences, reflecting in part aspects of the DNA that are recognized directly by the σ factor. These common features are often summarized in the form of a *consensus sequence* (**Figure 6–12**). A **consensus nucleotide sequence** is derived by comparing many sequences with the same basic function and tallying up the most common nucleotide found at each position. It

therefore serves as a summary or "average" of a large number of individual nucleotide sequences.

The DNA sequences of individual bacterial promoters differ in ways that determine their strength (the number of initiation events per unit time of the promoter). Evolutionary processes have fine-tuned each to initiate as often as necessary and have thereby created a wide spectrum of promoters. Promoters for genes that code for abundant proteins are much stronger than those associated with genes that encode rare proteins, and their nucleotide sequences are responsible for these differences.

Like bacterial promoters, transcription terminators also have a wide range of sequences, with the potential to form a simple hairpin RNA structure being the most important common feature. Since an almost unlimited number of nucleotide sequences have this potential, terminator sequences are even more heterogeneous than promoter sequences.

We have discussed bacterial promoters and terminators in some detail to illustrate an important point regarding the analysis of genome sequences. Although we know a great deal about bacterial promoters and terminators and can construct consensus sequences that summarize their most salient features, their variation in nucleotide sequence makes it difficult to definitively locate them simply by analysis of the nucleotide sequence of a genome. It is even more difficult to locate analogous sequences in eucaryotic genomes, due in part to the excess DNA carried in them. Often, we need additional information, some of it from direct experimentation, to locate and accurately interpret the short DNA signals contained in genomes.

Since DNA is double-stranded, two different RNA molecules could in principle be transcribed from any gene, using each of the two DNA strands as a template. However, a gene typically has only a single promoter, and because the promoter's nucleotide sequence is asymmetric (see Figure 6–12), the polymerase can bind in only one orientation. The polymerase synthesizes RNA in the 5′-to-3′ direction, and it can therefore only transcribe one strand per gene (**Figure 6–13**). Genome sequences reveal that the DNA strand used as the template for RNA synthesis varies from gene to gene depending on the location and orientation of the promoter (**Figure 6–14**).

Having considered transcription in bacteria, we now turn to the situation in eucaryotes, where the synthesis of RNA molecules is a much more elaborate affair.

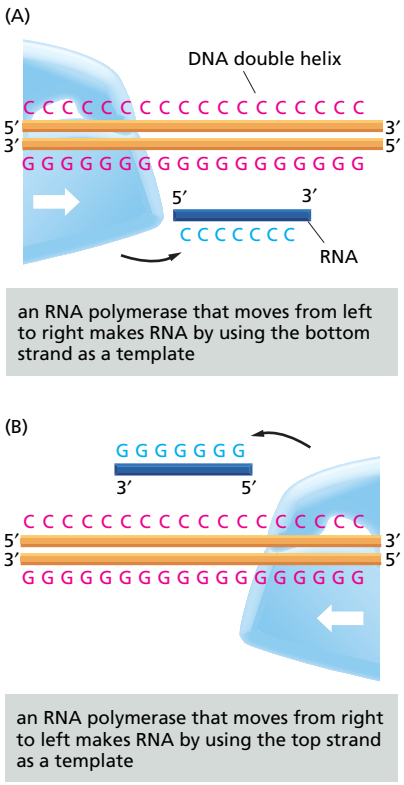


Figure 6–13 The importance of RNA polymerase orientation. The DNA strand serving as template must be traversed in a 3′-to-5′ direction. Thus, the direction of RNA polymerase movement determines which of the two DNA strands is to serve as a template for the synthesis of RNA, as shown in (A) and (B). Polymerase direction is, in turn, determined by the orientation of the promoter sequence, the site at which the RNA polymerase begins transcription.

Transcription Initiation in Eucaryotes Requires Many Proteins

In contrast to bacteria, which contain a single type of RNA polymerase, eucaryotic nuclei have three: *RNA polymerase I, RNA polymerase II,* and *RNA polymerase III.* The three polymerases are structurally similar to one another (and to the bacterial enzyme) and share some common subunits, but they transcribe different types of genes (**Table 6–2**). RNA polymerases I and III transcribe the genes encoding transfer RNA, ribosomal RNA, and various small RNAs. RNA polymerase II transcribes most genes, including all those that encode proteins, and our subsequent discussion therefore focuses on this enzyme.

Although eucaryotic RNA polymerase II has many structural similarities to bacterial RNA polymerase (**Figure 6–15**), there are several important differences in the way in which the bacterial and eucaryotic enzymes function, two of which concern us immediately.

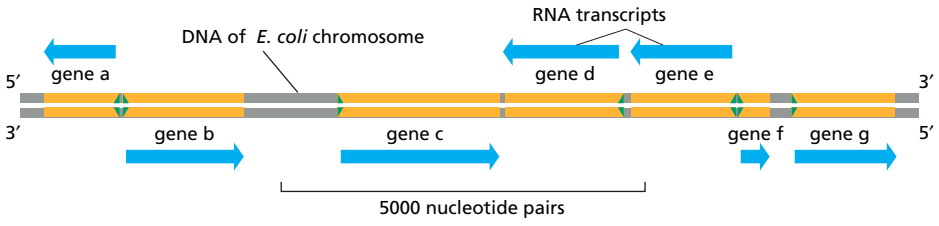


Figure 6–14 Directions of transcription along a short portion of a bacterial chromosome. Some genes are transcribed using one DNA strand as a template, while others are transcribed using the other DNA strand. The direction of transcription is determined by the promoter at the beginning of each gene *(green arrowheads).* This diagram shows approximately 0.2% (9000 base pairs) of the *E. coli* chromosome. The genes transcribed from *left* to *right* use the bottom DNA strand as the template; those transcribed from *right* to *left* use the top strand as the template.

Table 6–2 The Three RNA Polymerases in Eucaryotic Cells

TYPE OF POLYMERASE	GENES TRANSCRIBED
RNA polymerase I	5.8S, 18S, and 28S rRNA genes
RNA polymerase II	all protein-coding genes, plus snoRNA genes, miRNA genes, siRNA genes, and most snRNA genes
RNA polymerase III	tRNA genes, 5S rRNA genes, some snRNA genes and genes for other small RNAs

The rRNAs are named according to their "S" values, which refer to their rate of sedimentation in an ultracentrifuge. The larger the S value, the larger the rRNA.

1. While bacterial RNA polymerase requires only a single additional protein (σ factor) for transcription initiation to occur *in vitro*, eucaryotic RNA polymerases require many additional proteins, collectively called the *general transcription factors*.
2. Eucaryotic transcription initiation must deal with the packing of DNA into nucleosomes and higher-order forms of chromatin structure, features absent from bacterial chromosomes.

RNA Polymerase II Requires General Transcription Factors

The **general transcription factors** help to position eucaryotic RNA polymerase correctly at the promoter, aid in pulling apart the two strands of DNA to allow transcription to begin, and release RNA polymerase from the promoter into the elongation mode once transcription has begun. <CTAT> The proteins are "general" because they are needed at nearly all promoters used by RNA polymerase II; consisting of a set of interacting proteins, they are designated as *TFII* (for transcription factor for polymerase II), and are denoted arbitrarily as TFIIB, TFIID, and so on. In a broad sense, the eucaryotic general transcription factors carry out functions equivalent to those of the σ factor in bacteria; indeed, portions of TFIIF have the same three-dimensional structure as the equivalent portions of σ.

Figure 6–16 illustrates how the general transcription factors assemble at promoters used by RNA polymerase II, and **Table 6–3** summarizes their activities. The assembly process begins when the general transcription factor TFIID binds to a short double-helical DNA sequence primarily composed of T and A nucleotides. For this reason, this sequence is known as the TATA sequence, or **TATA box**, and the subunit of TFIID that recognizes it is called TBP (for TATA-binding protein). The TATA box is typically located 25 nucleotides upstream from the transcription start site. It is not the only DNA sequence that signals the start of transcription (**Figure 6–17**), but for most polymerase II promoters it is the most important. The binding of TFIID causes a large distortion in the DNA

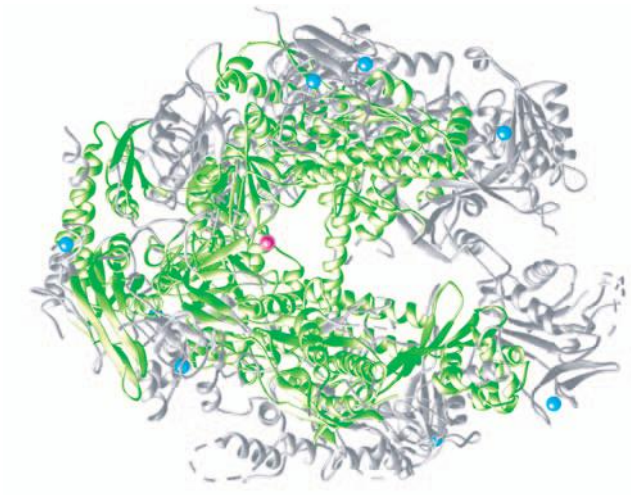


Figure 6–15 Structural similarity between a bacterial RNA polymerase and a eucaryotic RNA polymerase II. Regions of the two RNA polymerases that have similar structures are indicated in *green*. The eucaryotic polymerase is larger than the bacterial enzyme (12 subunits instead of 5), and some of the additional regions are shown in *gray*. The *blue* spheres represent Zn atoms that serve as structural components of the polymerases, and the *red* sphere represents the Mg atom present at the active site, where polymerization takes place. The RNA polymerases in all modern-day cells (bacteria, archaea, and eucaryotes) are closely related, indicating that the basic features of the enzyme were in place before the divergence of the three major branches of life. (Courtesy of P. Cramer and R. Kornberg.)

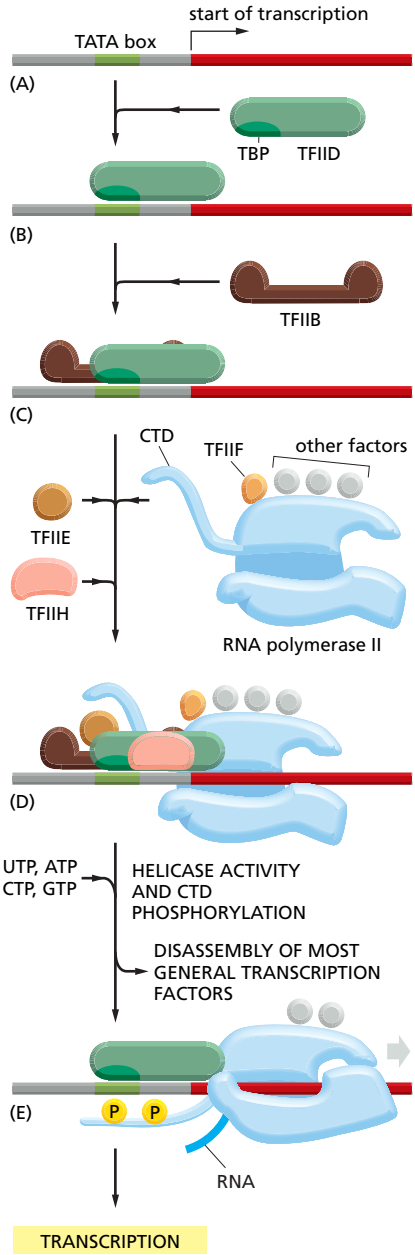


Figure 6–16 Initiation of transcription of a eucaryotic gene by RNA polymerase II. To begin transcription, RNA polymerase requires several general transcription factors. (A) The promoter contains a DNA sequence called the TATA box, which is located 25 nucleotides away from the site at which transcription is initiated. (B) Through its subunit TBP, TFIID recognizes and binds the TATA box, which then enables the adjacent binding of TFIIB (C). For simplicity the DNA distortion produced by the binding of TFIID (see Figure 6–18) is not shown. (D) The rest of the general transcription factors, as well as the RNA polymerase itself, assemble at the promoter. (E) TFIIH then uses ATP to pry apart the DNA double helix at the transcription start point, locally exposing the template strand. TFIIH also phosphorylates RNA polymerase II, changing its conformation so that the polymerase is released from the general factors and can begin the elongation phase of transcription. As shown, the site of phosphorylation is a long C-terminal polypeptide tail, also called the C-terminal domain (CTD), that extends from the polymerase molecule. The assembly scheme shown in the figure was deduced from experiments performed *in vitro*, and the exact order in which the general transcription factors assemble on promoters may vary from gene to gene *in vivo*. The general transcription factors have been highly conserved in evolution; some of those from human cells can be replaced in biochemical experiments by the corresponding factors from simple yeasts.

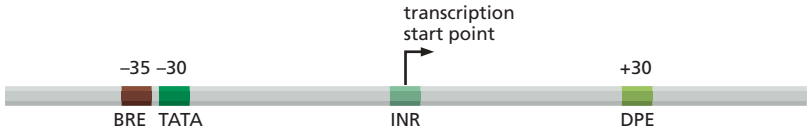
of the TATA box (**Figure 6–18**). This distortion is thought to serve as a physical landmark for the location of an active promoter in the midst of a very large genome, and it brings DNA sequences on both sides of the distortion together to allow for subsequent protein assembly steps. Other factors then assemble, along with RNA polymerase II, to form a complete *transcription initiation complex* (see Figure 6–16). The most complicated of the general transcription factors is TFIIH. Consisting of 9 subunits, it is nearly as large as RNA polymerase II itself and, as we shall see shortly, performs several enzymatic steps needed for the initiation of transcription.

After forming a transcription initiation complex on the promoter DNA, RNA polymerase II must gain access to the template strand at the transcription start point. TFIIH, which contains a DNA helicase as one of its subunits, makes this step possible by hydrolyzing ATP and unwinding the DNA, thereby exposing the template strand. Next, RNA polymerase II, like the bacterial polymerase, remains at the promoter synthesizing short lengths of RNA until it undergoes a series of conformational changes that allow it to move away from the promoter and enter the elongation phase of transcription. A key step in this transition is the addition of phosphate groups to the "tail" of the RNA polymerase (known as the CTD or C-terminal domain). In humans, the CTD consists of 52 tandem repeats of a seven-amino-acid sequence, which extend from the RNA polymerase core structure. During transcription initiation, the serine located at the

Table 6–3 The General Transcription Factors Needed for Transcription Initiation by Eucaryotic RNA Polymerase II

NAME	NUMBER OF SUBUNITS	ROLES IN TRANSITION INITIATION
TFIID		
TBP subunit	1	recognizes TATA box
TAF subunits	~11	recognizes other DNA sequences near the transcription start point; regulates DNA-binding by TBP
TFIIB	1	recognizes BRE element in promoters; accurately positions RNA polymerase at the start site of transcription
TFIIF	3	stabilizes RNA polymerase interaction with TBP and TFIIB; helps attract TFIIE and TFIIH
TFIIE	2	attracts and regulates TFIIH
TFIIH	9	unwinds DNA at the transcription start point, phosphorylates Ser5 of the RNA polymerase CTD; releases RNA polymerase from the promoter

TFIID is composed of TBP and ~11 additional subunits called TAFs (TBP-associated factors); CTD, C-terminal domain.



element	consensus sequence	general transcription factor
BRE	G/C G/C G/A C G C C	TFIIB
TATA	T A T A A/T A A/T	TBP
INR	C/T C/T A N T/A C/T C/T	TFIID
DPE	A/G G A/T C G T G	TFIID

Figure 6–17 Consensus sequences found in the vicinity of eucaryotic RNA polymerase II start points. The name given to each consensus sequence *(first column)* and the general transcription factor that recognizes it *(last column)* are indicated. N indicates any nucleotide, and two nucleotides separated by a slash indicate an equal probability of either nucleotide at the indicated position. In reality, each consensus sequence is a shorthand representation of a histogram similar to that of Figure 6–12.

For most RNA polymerase II transcription start points, only two or three of the four sequences are present. For example, many polymerase II promoters have a TATA box sequence, but those that do not typically have a "strong" INR sequence. Although most of the DNA sequences that influence transcription initiation are located upstream of the transcription start point, a few, such as the DPE shown in the figure, are located in the transcribed region.

fifth position in the repeat sequence (Ser5) is phosphorylated by TFIIH, which contains a protein kinase in another of its subunits (see Figure 6–16D and E). The polymerase can then disengage from the cluster of general transcription factors. During this process, it undergoes a series of conformational changes that tighten its interaction with DNA, and it acquires new proteins that allow it to transcribe for long distances, and in some cases for many hours, without dissociating from DNA.

Once the polymerase II has begun elongating the RNA transcript, most of the general transcription factors are released from the DNA so that they are available to initiate another round of transcription with a new RNA polymerase molecule. As we see shortly, the phosphorylation of the tail of RNA polymerase II also causes components of the RNA-processing machinery to load onto the polymerase and thus be positioned to modify the newly transcribed RNA as it emerges from the polymerase.

Polymerase II Also Requires Activator, Mediator, and Chromatin-Modifying Proteins

Studies of the behavior of RNA polymerase II and its general transcription factors on purified DNA templates *in vitro* established the model for transcription initiation just described. However, as discussed in Chapter 4, DNA in eucaryotic cells is packaged into nucleosomes, which are further arranged in higher-order

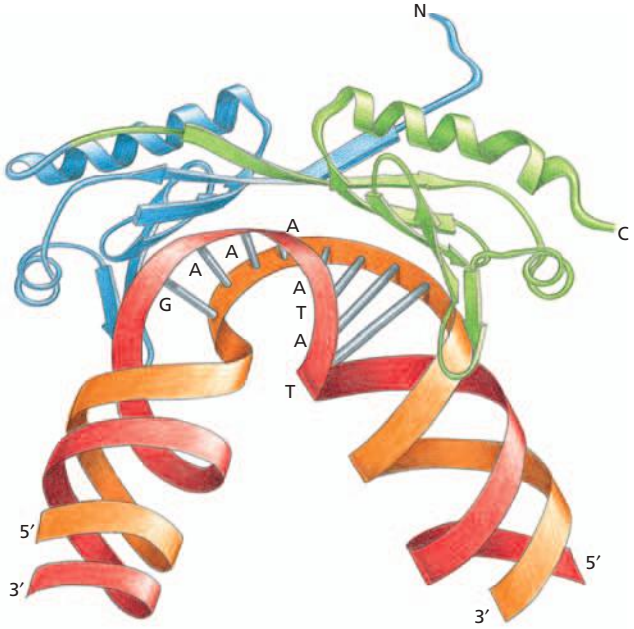


Figure 6–18 Three-dimensional structure of TBP (TATA-binding protein) bound to DNA. The TBP is the subunit of the general transcription factor TFIID that is responsible for recognizing and binding to the TATA box sequence in the DNA *(red)*. The unique DNA bending caused by TBP—two kinks in the double helix separated by partly unwound DNA—may serve as a landmark that helps to attract the other general transcription factors. TBP is a single polypeptide chain that is folded into two very similar domains *(blue* and *green)*. (Adapted from J.L. Kim et al., *Nature* 365:520–527, 1993. With permission from Macmillan Publishers Ltd.)

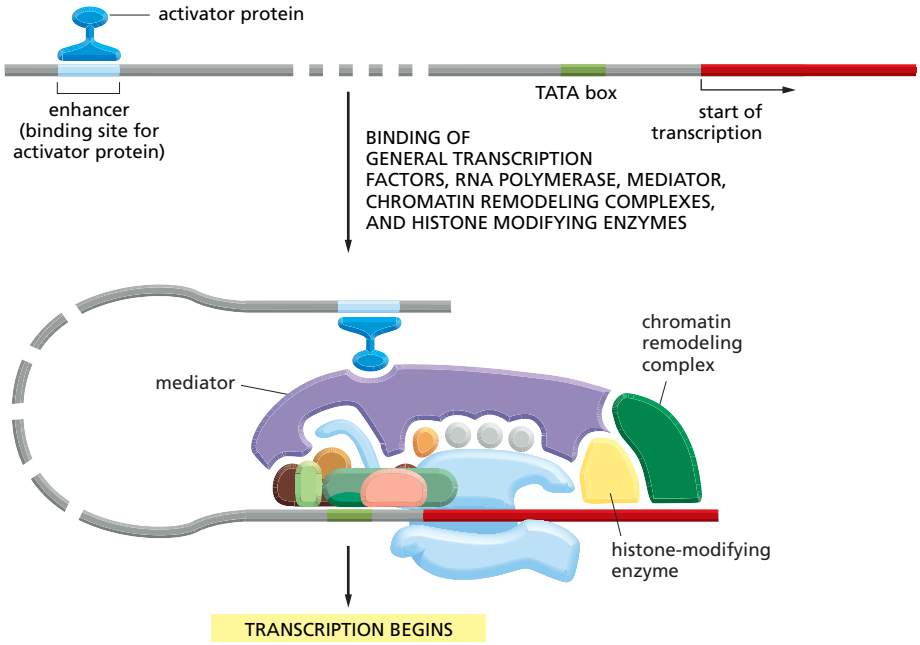


Figure 6–19 Transcription initiation by RNA polymerase II in a eucaryotic cell. Transcription initiation *in vivo* requires the presence of transcriptional activator proteins. As described in Chapter 7, these proteins bind to specific short sequences in DNA. Although only one is shown here, a typical eucaryotic gene has many activator proteins, which together determine its rate and pattern of transcription. Sometimes acting from a distance of several thousand nucleotide pairs (indicated by the dashed DNA molecule), these gene regulatory proteins help RNA polymerase, the general transcription factors, and the mediator all to assemble at the promoter. In addition, activators attract ATP-dependent chromatin remodeling complexes and histone acetylases.

As discussed in Chapter 4, the "default" state of chromatin is probably the 30-nm filament (see Figure 4–22), and this is likely to be a form of DNA upon which transcription is initiated. For simplicity, it is not shown in the figure.

chromatin structures. As a result, transcription initiation in a eucaryotic cell is more complex and requires even more proteins than it does on purified DNA. First, gene regulatory proteins known as *transcriptional activators* must bind to specific sequences in DNA and help to attract RNA polymerase II to the start point of transcription (**Figure 6–19**). We discuss the role of activators in Chapter 7, because they are one of the main ways in which cells regulate expression of their genes. Here we simply note that their presence on DNA is required for transcription initiation in a eucaryotic cell. Second, eucaryotic transcription initiation *in vivo* requires the presence of a protein complex known as *Mediator*, which allows the activator proteins to communicate properly with the polymerase II and with the general transcription factors. Finally, transcription initiation in a eucaryotic cell typically requires the local recruitment of chromatin-modifying enzymes, including chromatin remodeling complexes and histone-modifying enzymes. As discussed in Chapter 4, both types of enzymes can allow greater access to the DNA present in chromatin, and by doing so, they facilitate the assembly of the transcription initiation machinery onto DNA. We will revisit the role of these enzymes in transcription initiation in Chapter 7.

As illustrated in Figure 6–19, many proteins (well over 100 individual subunits) must assemble at the start point of transcription to initiate transcription in a eucaryotic cell. The order of assembly of these proteins does not seem to follow a prescribed pathway; rather, the order differs from gene to gene. Indeed, some of these different protein complexes may interact with each other away from the DNA and be brought to DNA as preformed subassemblies. To begin transcribing, RNA polymerase II must be released from this large complex of proteins, and, in addition to the steps described in Figure 6–16, this often requires the *in situ* proteolysis of the activator protein. We return to some of these issues in Chapter 7, where we discuss how eucaryotic cells can regulate the process of transcription initiation.

Transcription Elongation Produces Superhelical Tension in DNA

Once it has initiated transcription, RNA polymerase does not proceed smoothly along a DNA molecule; rather, it moves jerkily, pausing at some sequences and rapidly transcribing through others. Elongating RNA polymerases, both bacterial and eucaryotic, are associated with a series of *elongation factors*, proteins that decrease the likelihood that RNA polymerase will dissociate before it reaches the end of a gene. These factors typically associate with RNA polymerase

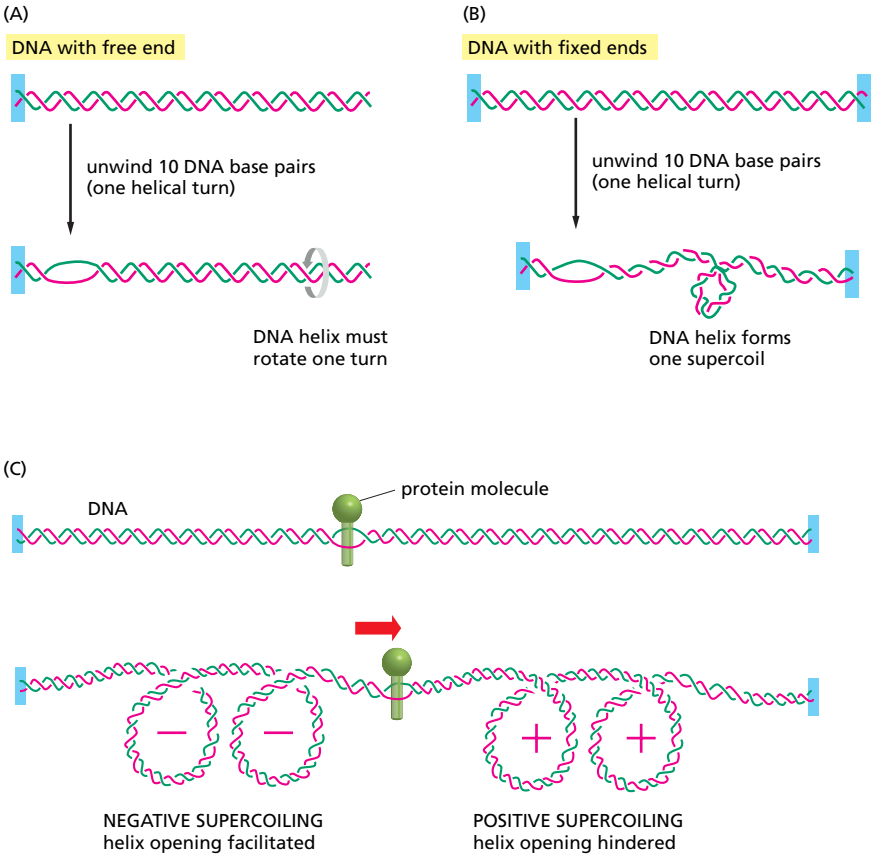


Figure 6–20 Superhelical tension in DNA causes DNA supercoiling. (A) For a DNA molecule with one free end (or a nick in one strand that serves as a swivel), the DNA double helix rotates by one turn for every 10 nucleotide pairs opened. (B) If rotation is prevented, superhelical tension is introduced into the DNA by helix opening. One way of accommodating this tension would be to increase the helical twist from 10 to 11 nucleotide pairs per turn in the double helix that remains; the DNA helix, however, resists such a deformation in a springlike fashion, preferring to relieve the superhelical tension by bending into supercoiled loops. As a result, one DNA supercoil forms in the DNA double helix for every 10 nucleotide pairs opened. The supercoil formed in this case is a positive supercoil. (C) Supercoiling of DNA is induced by a protein tracking through the DNA double helix. The two ends of the DNA shown here are unable to rotate freely relative to each other, and the protein molecule is assumed also to be prevented from rotating freely as it moves. Under these conditions, the movement of the protein causes an excess of helical turns to accumulate in the DNA helix ahead of the protein and a deficit of helical turns to arise in the DNA behind the protein, as shown.

shortly after initiation and help polymerases to move through the wide variety of different DNA sequences that are found in genes. Eucaryotic RNA polymerases must also contend with chromatin structure as they move along a DNA template, and they are typically aided by ATP-dependent chromatin remodeling complexes (see pp. 215–216). These complexes may move with the polymerase or may simply seek out and rescue the occasional stalled polymerase. In addition, some elongation factors associated with eucaryotic RNA polymerase facilitate transcription through nucleosomes without requiring additional energy. It is not yet understood in detail how this is accomplished, but these proteins can transiently dislodge H2A–H2B dimers from the nucleosome core, replacing them as the polymerase moves through the nucleosome.

There is yet another barrier to elongating polymerases, both bacterial and eucaryotic. To discuss this issue, we need first to consider a subtle property inherent in the DNA double helix called **DNA supercoiling**. DNA supercoiling represents a conformation that DNA adopts in response to superhelical tension; conversely, creating various loops or coils in the helix can create such tension. **Figure 6–20** illustrates the topological constraints that cause DNA supercoiling. There are approximately 10 nucleotide pairs for every helical turn in a DNA double helix. Imagine a helix whose two ends are fixed with respect to each other (as they are in a DNA circle, such as a bacterial chromosome, or in a tightly clamped loop, as is thought to exist in eucaryotic chromosomes). In this case, one large DNA supercoil will form to compensate for each 10 nucleotide pairs that are opened (unwound). The formation of this supercoil is energetically favorable because it restores a normal helical twist to the base-paired regions that remain, which would otherwise need to be overwound because of the fixed ends.

RNA polymerase also creates superhelical tension as it moves along a stretch of DNA that is anchored at its ends (see Figure 6–20C). As long as the polymerase is not free to rotate rapidly (and such rotation is unlikely given the size of RNA polymerases and their attached transcripts), a moving polymerase generates

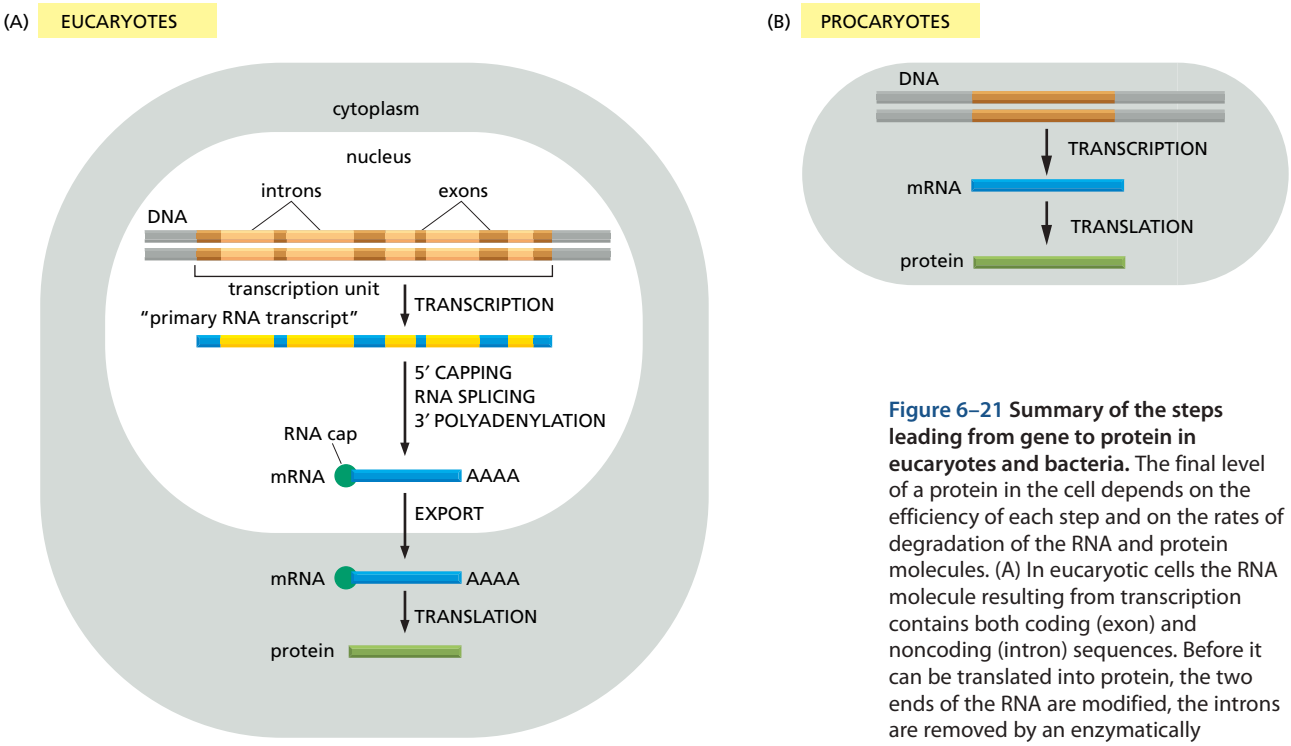


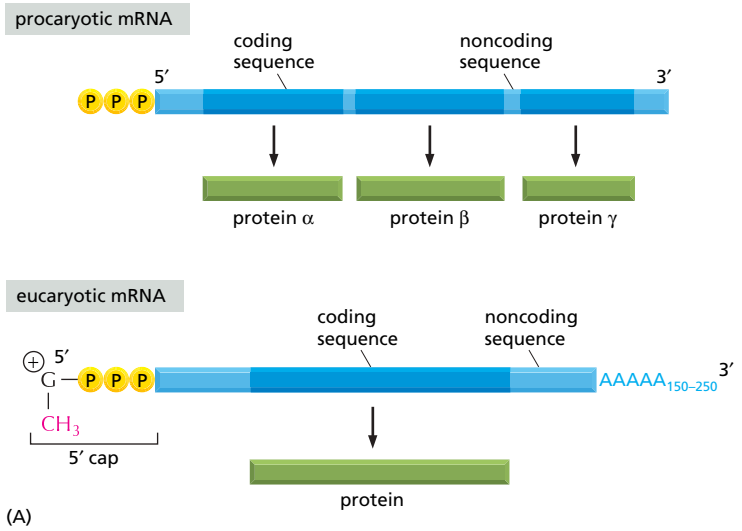
Figure 6–21 Summary of the steps leading from gene to protein in eucaryotes and bacteria. The final level of a protein in the cell depends on the efficiency of each step and on the rates of degradation of the RNA and protein molecules. (A) In eucaryotic cells the RNA molecule resulting from transcription contains both coding (exon) and noncoding (intron) sequences. Before it can be translated into protein, the two ends of the RNA are modified, the introns are removed by an enzymatically catalyzed RNA splicing reaction, and the resulting mRNA is transported from the nucleus to the cytoplasm. Although the steps in this figure are depicted as occurring one at a time, in a sequence, in reality they can occur concurrently. For example, the RNA cap is added and splicing typically begins before transcription has been completed. Because of the coupling between transcription and RNA processing, primary transcripts—the RNAs that would, in theory, be produced if no processing had occurred—are found only rarely. (B) In procaryotes the production of mRNA is much simpler. The 5′ end of an mRNA molecule is produced by the initiation of transcription, and the 3′ end is produced by the termination of transcription. Since procaryotic cells lack a nucleus, transcription and translation take place in a common compartment. In fact, the translation of a bacterial mRNA often begins before its synthesis has been completed.

positive superhelical tension in the DNA in front of it and negative helical tension behind it. For eucaryotes, this situation is thought to provide a bonus: the positive superhelical tension ahead of the polymerase makes the DNA helix more difficult to open, but this tension should facilitate the unwrapping of DNA in nucleosomes, as the release of DNA from the histone core helps to relax positive superhelical tension.

Any protein that propels itself alone along a DNA strand of a double helix tends to generate superhelical tension. In eucaryotes, DNA topoisomerase enzymes rapidly remove this superhelical tension (see p. 278). But in bacteria a specialized topoisomerase called *DNA gyrase* uses the energy of ATP hydrolysis to pump supercoils continuously into the DNA, thereby maintaining the DNA under constant tension. These are *negative supercoils,* having the opposite handedness from the *positive supercoils* that form when a region of DNA helix opens (see Figure 6–20B). Whenever a region of helix opens, it removes these negative supercoils from bacterial DNA, reducing the superhelical tension. DNA gyrase therefore makes the opening of the DNA helix in bacteria energetically favorable compared with helix opening in DNA that is not supercoiled. For this reason, it usually facilitates those genetic processes in bacteria, including the initiation of transcription by bacterial RNA polymerase, that require helix opening (see Figure 6–11).

Transcription Elongation in Eucaryotes Is Tightly Coupled to RNA Processing

We have seen that bacterial mRNAs are synthesized solely by the RNA polymerase starting and stopping at specific spots on the genome. The situation in eucaryotes is substantially different. In particular, transcription is only the first of several steps needed to produce an mRNA. Other critical steps are the covalent modification of the ends of the RNA and the removal of *intron sequences* that are discarded from the middle of the RNA transcript by the process of *RNA splicing* (**Figure 6–21**).



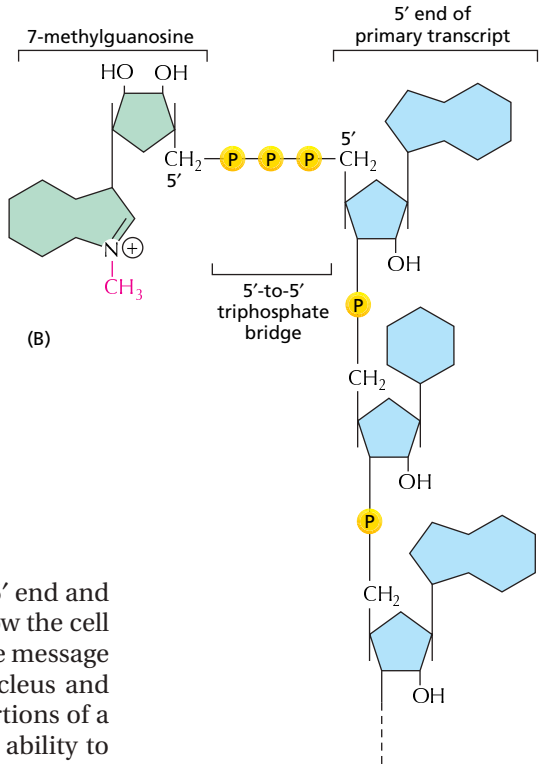


Figure 6–22 A comparison of the structures of procaryotic and eucaryotic mRNA molecules. (A) The 5′ and 3′ ends of a bacterial mRNA are the unmodified ends of the chain synthesized by the RNA polymerase, which initiates and terminates transcription at those points, respectively. The corresponding ends of a eucaryotic mRNA are formed by adding a 5′ cap and by cleavage of the pre-mRNA transcript and the addition of a poly-A tail, respectively. The figure also illustrates another difference between the procaryotic and eucaryotic mRNAs: bacterial mRNAs can contain the instructions for several different proteins, whereas eucaryotic mRNAs nearly always contain the information for only a single protein. (B) The structure of the cap at the 5′ end of eucaryotic mRNA molecules. Note the unusual 5′-to-5′ linkage of the 7-methyl G to the remainder of the RNA. Many eucaryotic mRNAs carry an additional modification: the 2′-hydroxyl group on the second ribose sugar in the mRNA is methylated (not shown).

Both ends of eucaryotic mRNAs are modified: by *capping* on the 5′ end and by *polyadenylation* of the 3′ end (**Figure 6–22**). These special ends allow the cell to assess whether both ends of an mRNA molecule are present (and the message is therefore intact) before it exports the RNA sequence from the nucleus and translates it into protein. RNA splicing joins together the different portions of a protein coding sequence, and it provides higher eucaryotes with the ability to synthesize several different proteins from the same gene.

An ingenious mechanism couples all of the above RNA processing steps to transcription elongation. As discussed previously, a key step in transcription initiation by RNA polymerase II is the phosphorylation of the RNA polymerase II tail, called the CTD (C-terminal domain). This phosphorylation proceeds gradually as the RNA polymerase initiates transcription and moves along the DNA. It not only helps dissociate the RNA polymerase II from other proteins present at the start point of transcription, but also allows a new set of proteins to associate with the RNA polymerase tail that function in transcription elongation and RNA processing. As discussed next, some of these processing proteins seem to "hop" from the polymerase tail onto the nascent RNA molecule to begin processing it as it emerges from the RNA polymerase. Thus, we can view RNA polymerase II in its elongation mode as an RNA factory that both transcribes DNA into RNA and processes the RNA it produces (**Figure 6–23**). Fully extended, the CTD is nearly 10 times longer than the remainder of RNA polymerase and, in effect, it serves as a tether, holding a variety of proteins close by until they are needed. This strategy, which speeds up the rate of subsequent reactions, is one commonly observed in the cell (see Figures 4–69 and 16–38).

RNA Capping Is the First Modification of Eucaryotic Pre-mRNAs

As soon as RNA polymerase II has produced about 25 nucleotides of RNA, the 5′ end of the new RNA molecule is modified by addition of a cap that consists of a modified guanine nucleotide (see Figure 6–22B). Three enzymes, acting in succession, perform the capping reaction: one (a phosphatase) removes a phosphate from the 5′ end of the nascent RNA, another (a guanyl transferase) adds a GMP in a reverse linkage (5′ to 5′ instead of 5′ to 3′), and a third (a methyl transferase) adds a methyl group to the guanosine (**Figure 6–24**). Because all three enzymes bind to the RNA polymerase tail phosphorylated at serine-5 position, the modification added by TFIIH during transcription initiation, they are poised to modify the 5′ end of the nascent transcript as soon as it emerges from the polymerase.

The 5′-methyl cap signifies the 5′ end of eucaryotic mRNAs, and this landmark helps the cell to distinguish mRNAs from the other types of RNA molecules present in the cell. For example, RNA polymerases I and III produce uncapped

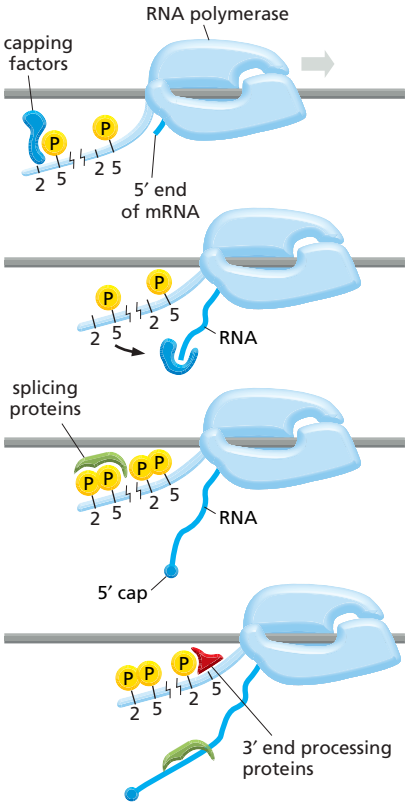


Figure 6–23 Eucaryotic RNA polymerase II as an "RNA factory." As the polymerase transcribes DNA into RNA, it carries pre-mRNA-processing proteins on its tail that are transferred to the nascent RNA at the appropriate time. The tail, known as the CTD, contains 52 tandem repeats of a seven amino acid sequence, and there are two serines in each repeat. The capping proteins first bind to the RNA polymerase tail when it is phosphorylated on Ser5 of the heptad repeat late in the process of transcription initiation (see Figure 6–16). This strategy ensures that the RNA molecule is efficiently capped as soon as its 5′ end emerges from the RNA polymerase. As the polymerase continues transcribing, its tail is extensively phosphorylated on the Ser2 positions by a kinase associated with the elongating polymerase and is eventually dephosphorylated at Ser5 positions. These further modifications attract splicing and 3′-end processing proteins to the moving polymerase, positioning them to act on the newly synthesized RNA as it emerges from the RNA polymerase. There are many RNA-processing enzymes, and not all travel with the polymerase. For RNA splicing, for example, the tail carries only a few critical components; once transferred to an RNA molecule, they serve as a nucleation site for the remaining components.

When RNA polymerase II finishes transcribing a gene, it is released from DNA, soluble phosphatases remove the phosphates on its tail, and it can reinitiate transcription. Only the dephosphorylated form of RNA polymerase II is competent to begin RNA synthesis at a promoter.

RNAs during transcription, in part because these polymerases lack a CTD. In the nucleus, the cap binds a protein complex called CBC (cap-binding complex), which, as we discuss in subsequent sections, helps the RNA to be properly processed and exported. The 5′-methyl cap also has an important role in the translation of mRNAs in the cytosol, as we discuss later in the chapter.

RNA Splicing Removes Intron Sequences from Newly Transcribed Pre-mRNAs

As discussed in Chapter 4, the protein coding sequences of eucaryotic genes are typically interrupted by noncoding intervening sequences (introns). Discovered in 1977, this feature of eucaryotic genes came as a surprise to scientists, who had been, until that time, familiar only with bacterial genes, which typically consist of a continuous stretch of coding DNA that is directly transcribed into mRNA. In marked contrast, eucaryotic genes were found to be broken up into small pieces of coding sequence (*expressed sequences* or **exons**) interspersed with much longer *intervening sequences* or **introns**; thus, the coding portion of a eucaryotic gene is often only a small fraction of the length of the gene (**Figure 6–25**).

Both intron and exon sequences are transcribed into RNA. The intron sequences are removed from the newly synthesized RNA through the process of **RNA splicing**. The vast majority of RNA splicing that takes place in cells functions in the production of mRNA, and our discussion of splicing focuses on this so-called precursor-mRNA (or pre-mRNA) splicing. Only after 5′ and 3′ end processing and splicing have taken place is such RNA termed mRNA.

Each splicing event removes one intron, proceeding through two sequential phosphoryl-transfer reactions known as transesterifications; these join two exons while removing the intron as a "lariat" (**Figure 6–26**). Since the number of high-energy phosphate bonds remains the same, these reactions could in

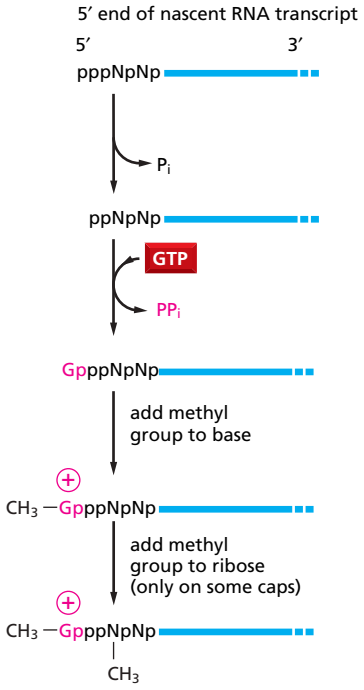


Figure 6–24 The reactions that cap the 5′ end of each RNA molecule synthesized by RNA polymerase II. The final cap contains a novel 5′-to-5′ linkage between the positively charged 7-methyl G residue and the 5′ end of the RNA transcript (see Figure 6–22B). The letter N represents any one of the four ribonucleotides, although the nucleotide that starts an RNA chain is usually a purine (an A or a G). (After A.J. Shatkin, *BioEssays* 7:275–277, 1987. With permission from ICSU Press.)

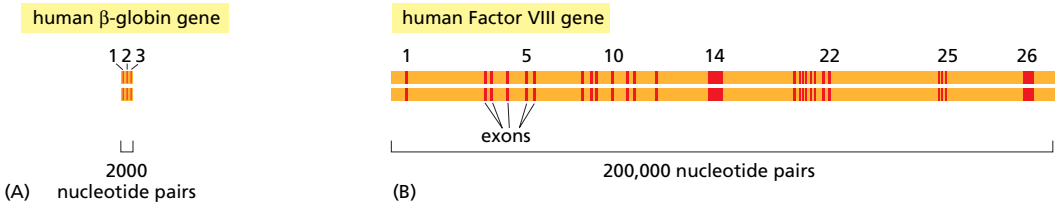


Figure 6–25 Structure of two human genes showing the arrangement of exons and introns. (A) The relatively small β-globin gene, which encodes one of the subunits of the oxygen-carrying protein hemoglobin, contains 3 exons (see also Figure 4–7). (B) The much larger Factor VIII gene contains 26 exons; it codes for a protein (Factor VIII) that functions in the blood-clotting pathway. The most prevalent form of hemophilia results from mutations in this gene.

principle take place without nucleoside triphosphate hydrolysis. However, the machinery that catalyzes pre-mRNA splicing is complex, consisting of 5 additional RNA molecules and as many as 200 proteins, and it hydrolyzes many ATP molecules per splicing event. This additional complexity ensures that splicing is accurate, while at the same time being flexible enough to deal with the enormous variety of introns found in a typical eucaryotic cell.

It may seem wasteful to remove large numbers of introns by RNA splicing. In attempting to explain why it occurs, scientists have pointed out that the exon–intron arrangement would seem to facilitate the emergence of new and useful proteins over evolutionary time scales. Thus, the presence of numerous introns in DNA allows genetic recombination to readily combine the exons of different genes (see p. 140), enabling genes for new proteins to evolve more easily by the combination of parts of preexisting genes. The observation, described in Chapter 3, that many proteins in present-day cells resemble patchworks composed from a common set of protein *domains*, supports this idea.

RNA splicing also has a present-day advantage. The transcripts of many eucaryotic genes (estimated at 75% of genes in humans) are spliced in more than one way, thereby allowing the same gene to produce a corresponding set of different proteins (**Figure 6–27**). Rather than being the wasteful process it may have seemed at first sight, RNA splicing enables eucaryotes to increase the already enormous coding potential of their genomes. We shall return to this idea again in this chapter and the next, but we first need to describe the cellular machinery that performs this remarkable task.

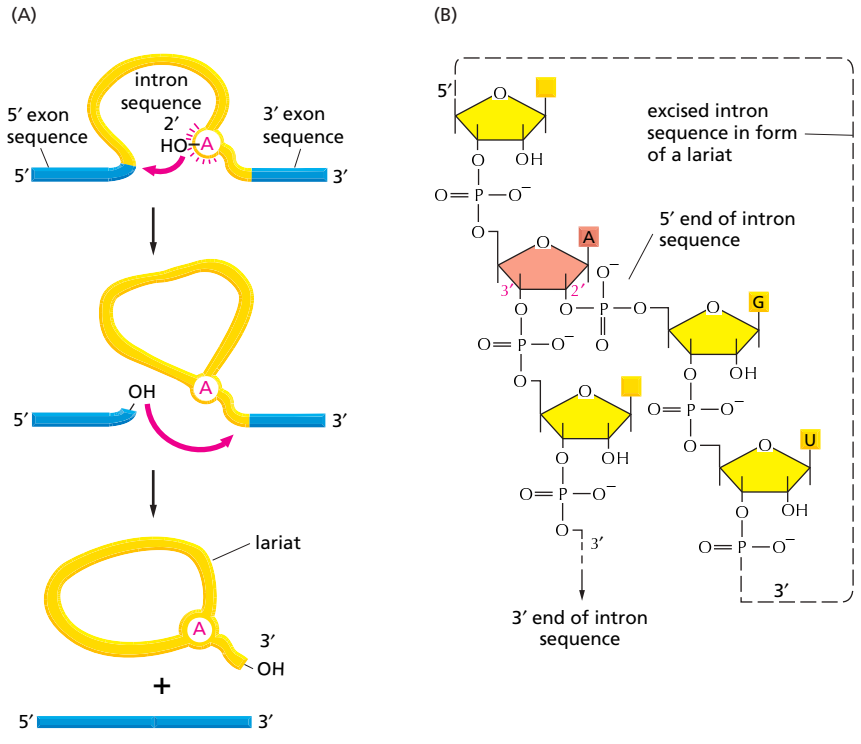


Figure 6–26 The pre-mRNA splicing reaction. (A) In the first step, a specific adenine nucleotide in the intron sequence (indicated in *red)* attacks the 5′ splice site and cuts the sugar-phosphate backbone of the RNA at this point. The cut 5′ end of the intron becomes covalently linked to the adenine nucleotide, as shown in detail in (B), thereby creating a loop in the RNA molecule. The released free 3′-OH end of the exon sequence then reacts with the start of the next exon sequence, joining the two exons together and releasing the intron sequence in the shape of a *lariat*. The two exon sequences thereby become joined into a continuous coding sequence; the released intron sequence is eventually degraded.

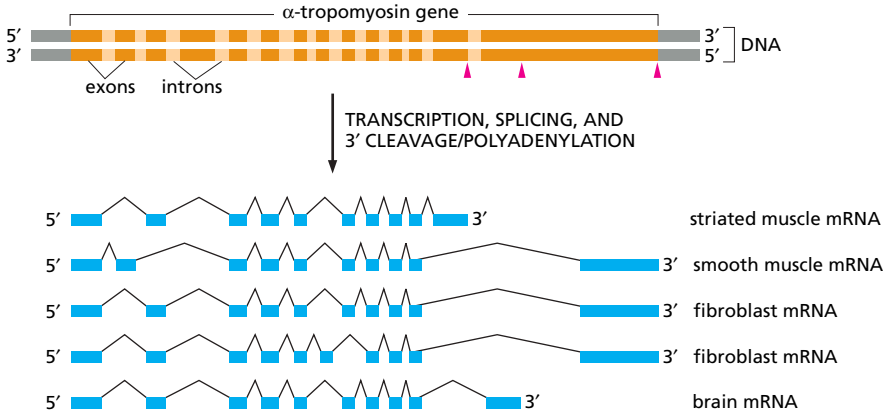


Figure 6–27 Alternative splicing of the α-tropomyosin gene from rat. α-Tropomyosin is a coiled-coil protein (see Figure 3–9) that regulates contraction in muscle cells. The primary transcript can be spliced in different ways, as indicated in the figure, to produce distinct mRNAs, which then give rise to variant proteins. Some of the splicing patterns are specific for certain types of cells. For example, the α-tropomyosin made in striated muscle is different from that made from the same gene in smooth muscle. The arrowheads in the top part of the figure mark the sites where cleavage and poly-A addition form the 3′ ends of the mature mRNAs.

Nucleotide Sequences Signal Where Splicing Occurs

The mechanism of pre-mRNA splicing shown in Figure 6–26 implies that the splicing machinery must recognize three portions of the precursor RNA molecule: the 5′ splice site, the 3′ splice site, and the branch point in the intron sequence that forms the base of the excised lariat. Not surprisingly, each site has a consensus nucleotide sequence that is similar from intron to intron and provides the cell with cues for where splicing is to take place (**Figure 6–28**). However, these consensus sequences are relatively short and can accommodate a high degree of sequence variability; as we shall see shortly, the cell incorporates additional types of information to ultimately choose exactly where, on each RNA molecule, splicing is to take place.

The high variability of the splicing consensus sequences presents a special challenge for scientists attempting to decipher genome sequences. Introns range in size from about 10 nucleotides to over 100,000 nucleotides, and choosing the precise borders of each intron is a difficult task even with the aid of powerful computers. The possibility of alternative splicing compounds the problem of predicting protein sequences solely from a genome sequence. This difficulty is one of the main barriers to identifying all of the genes in a complete genome sequence, and it is one of the primary reasons why we know only the approximate number of genes in the human genome.

RNA Splicing Is Performed by the Spliceosome

Unlike the other steps of mRNA production we have discussed, key steps in RNA splicing are performed by RNA molecules rather than proteins. Specialized RNA molecules recognize the nucleotide sequences that specify where splicing is to occur and also participate in the chemistry of splicing. These RNA molecules are relatively short (less than 200 nucleotides each), and there are five of them (U1, U2, U4, U5, and U6) involved in the major form of pre-mRNA splicing. Known as

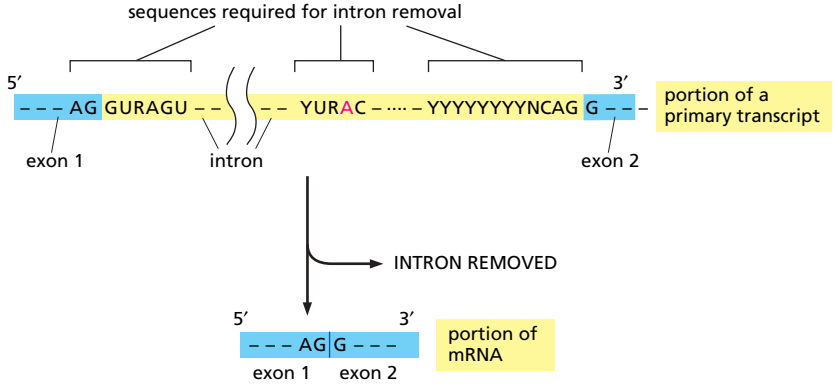
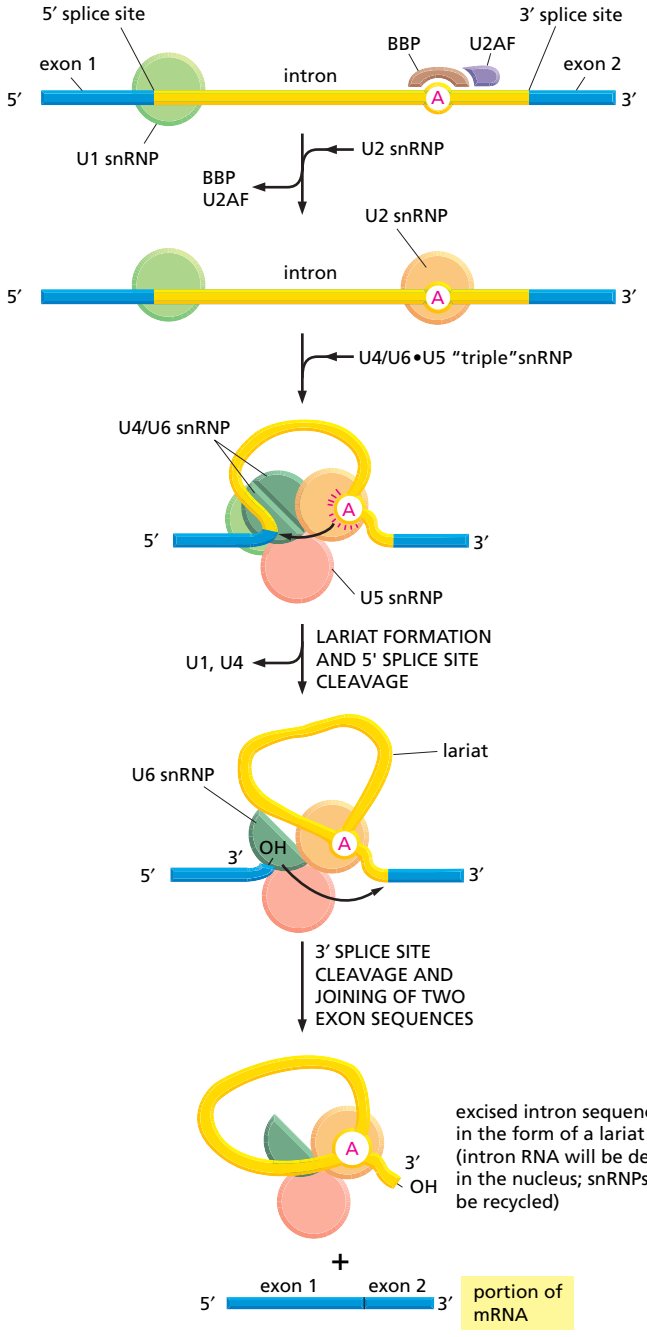


Figure 6–28 The consensus nucleotide sequences in an RNA molecule that signal the beginning and the end of most introns in humans. Only the three blocks of nucleotide sequences shown are required to remove an intron sequence; the rest of the intron can be occupied by any nucleotides. Here A, G, U, and C are the standard RNA nucleotides; R stands for purines (A or G); and Y stands for pyrimidines (C or U). The A highlighted in *red* forms the branch point of the lariat produced by splicing. Only the GU at the start of the intron and the AG at its end are invariant nucleotides in the splicing consensus sequences. Several different nucleotides can occupy the remaining positions (even the branch point A), although the indicated nucleotides are preferred. The distances along the RNA between the three splicing consensus sequences are highly variable; however, the distance between the branch point and 3′ splice junction is typically much shorter than that between the 5′ splice junction and the branch point.

snRNAs (**small nuclear RNAs**), each is complexed with at least seven protein subunits to form a snRNP (small nuclear ribonucleoprotein). These snRNPs form the core of the **spliceosome**, the large assembly of RNA and protein molecules that performs pre-mRNA splicing in the cell.

The spliceosome is a complex and dynamic machine. When studied *in vitro,* a few components of the spliceosome assemble on pre-mRNA and, as the splicing reaction proceeds, new components enter as those that have already performed their tasks are jettisoned (**Figure 6–29**). However, many scientists believe that, inside the cell, the spliceosome is a preexisting, loose assembly of all the components—capturing, splicing and releasing RNA as a coordinated unit, and undergoing extensive rearrangements each time a splice is made. During the splicing reaction, recognition of the 5′ splice junction, the branch-point site, and the 3′ splice junction is performed largely through base-pairing between the snRNAs and the consensus RNA sequences in the pre-mRNA substrate (**Figure**



The U1 snRNP forms base pairs with the 5′ splice junction (see Figure 6–30A) and the BBP (branch-point binding protein) and U2AF (U2 auxilliary factor) recognize the branch-point site.

The U2 snRNP displaces BBP and U2AF and forms base pairs with the branch-point site consensus sequence (see Figure 6–30B).

The U4/U6•U5 "triple" snRNP enters the reaction. In this triple snRNP, the U4 and U6 snRNAs are held firmly together by base-pair interactions. Subsequent rearrangements create the active site of the spliceosome and position the appropriate portions of the pre-mRNA substrate for the first phosphoryl-transferase reaction.

Several more RNA–RNA rearrangements occur that break apart the U4/U6 base pairs and allow the U6 snRNP to displace U1 at the 5′ splice junction (see Figure 6–30A) to form the active site for the second phosphoryl-transferase reaction, which completes the splice.

Figure 6–29 The pre-mRNA splicing mechanism. RNA splicing is catalyzed by an assembly of snRNPs (shown as *colored circles)* plus other proteins (most of which are not shown), which together constitute the spliceosome. The spliceosome recognizes the splicing signals on a pre-mRNA molecule, brings the two ends of the intron together, and provides the enzymatic activity for the two reaction steps (see Figure 6–26).

6–30). In the course of splicing, the spliceosome undergoes several shifts in which one set of base-pair interactions is broken and another is formed in its place. For example, U1 is replaced by U6 at the 5′ splice junction (see Figure 6–30A). This type of RNA–RNA rearrangement (in which the formation of one RNA–RNA interaction requires the disruption of another) occurs several times during the splicing reaction. It permits the checking and rechecking of RNA sequences before the chemical reaction is allowed to proceed, thereby increasing the accuracy of splicing.

The Spliceosome Uses ATP Hydrolysis to Produce a Complex Series of RNA–RNA Rearrangements

Although ATP hydrolysis is not required for the chemistry of RNA splicing *per se*, it is required for the assembly and rearrangements of the spliceosome. Some of the additional proteins that make up the spliceosome use the energy of ATP hydrolysis to break existing RNA–RNA interactions to allow the formation of new ones. In fact, all the steps shown previously in Figure 6–29—except the association of BBP with the branch-point site and U1 snRNP with the 5′ splice site—require ATP hydrolysis and additional proteins. Each successful splice requires as many as 200 proteins, if we include those that form the snRNPs.

The ATP-requiring RNA–RNA rearrangements that take place in the spliceosome occur within the snRNPs themselves and between the snRNPs and the pre-mRNA substrate. One of the most important functions of these rearrangements is the creation of the active catalytic site of the spliceosome. The strategy of creating an active site only after the assembly and rearrangement of splicing components on a pre-mRNA substrate is a particularly effective way to prevent wayward splicing.

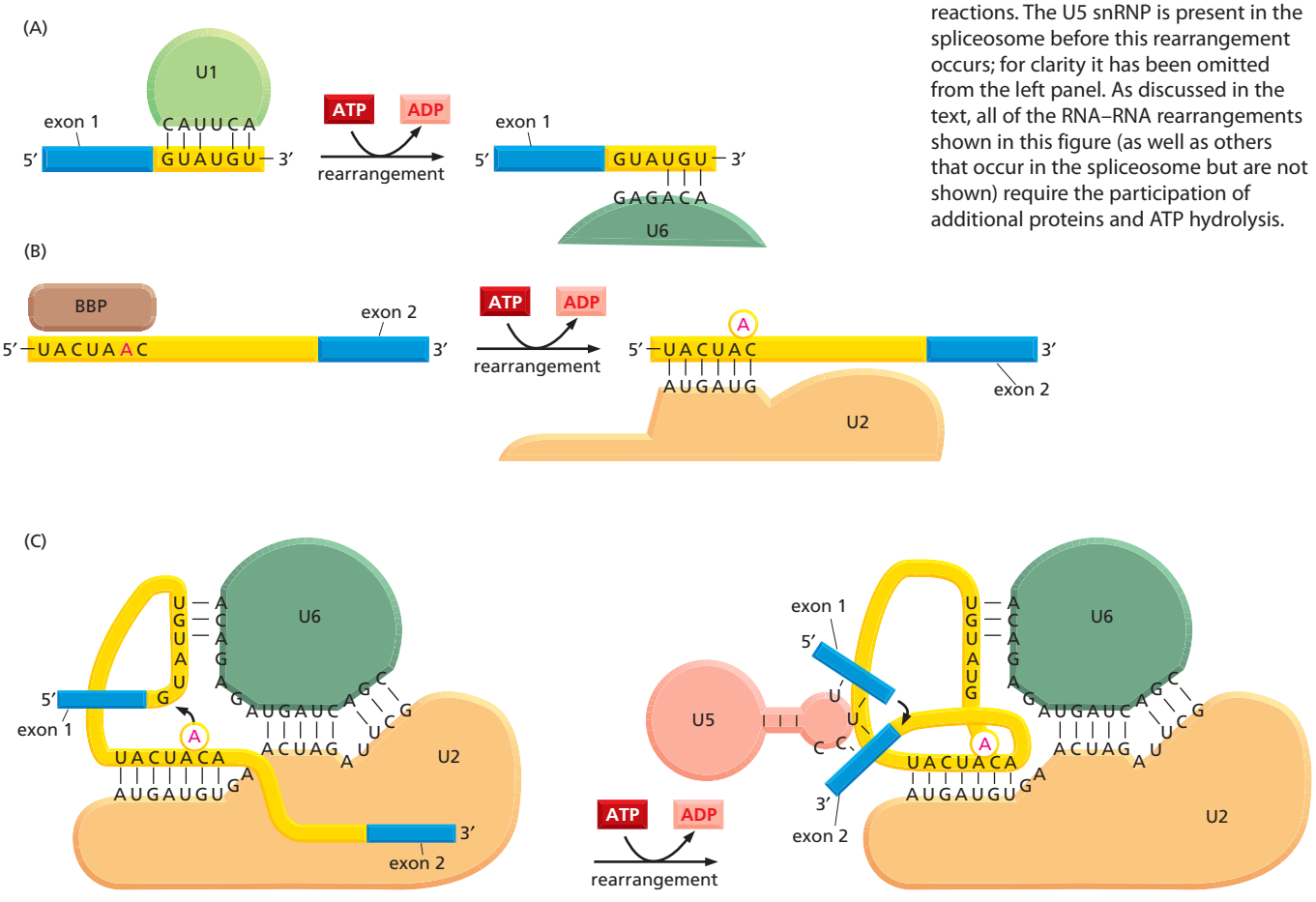


Figure 6–30 Several of the rearrangements that take place in the spliceosome during pre-mRNA splicing. Shown here are the details for the yeast *Saccharomyces cerevisiae*, in which the nucleotide sequences involved are slightly different from those in human cells. (A) The exchange of U1 snRNP for U6 snRNP occurs before the first phosphoryl-transfer reaction (see Figure 6–29). This exchange requires the 5′ splice site to be read by two different snRNPs, thereby increasing the accuracy of 5′ splice site selection by the spliceosome. (B) The branch-point site is first recognized by BBP and subsequently by U2 snRNP; as in (A), this "check and recheck" strategy provides increased accuracy of site selection. The binding of U2 to the branch point forces the appropriate adenine (in *red)* to be unpaired and thereby activates it for the attack on the 5′ splice site (see Figure 6–29). This, in combination with recognition by BBP, is the way in which the spliceosome accurately chooses the adenine that is ultimately to form the branch point. (C) After the first phosphoryl-transfer reaction *(left)* has occurred, a series of rearrangements brings the two exons into close proximity for the second phosphoryl-transfer reaction *(right)*. The snRNAs both position the reactants and provide (either all or in part) the catalytic sites for the two reactions. The U5 snRNP is present in the spliceosome before this rearrangement occurs; for clarity it has been omitted from the left panel. As discussed in the text, all of the RNA–RNA rearrangements shown in this figure (as well as others that occur in the spliceosome but are not shown) require the participation of additional proteins and ATP hydrolysis.

Perhaps the most surprising feature of the spliceosome is the nature of the catalytic site itself: it is largely (if not exclusively) formed by RNA molecules instead of proteins. In the last section of this chapter we discuss in general terms the structural and chemical properties of RNA that allow it to perform catalysis; here we need only consider that the U2 and U6 snRNAs in the spliceosome form a precise three-dimensional RNA structure that juxtaposes the 5′ splice site of the pre-mRNA with the branch-point site and probably performs the first transesterification reaction (see Figure 6–30C). In a similar way, the 5′ and 3′ splice junctions are brought together (an event requiring the U5 snRNA) to facilitate the second transesterification.

Once the splicing chemistry is completed, the snRNPs remain bound to the lariat. The disassembly of these snRNPs from the lariat (and from each other) requires another series of RNA–RNA rearrangements that require ATP hydrolysis, thereby returning the snRNAs to their original configuration so that they can be used again in a new reaction. At the completion of a splice, the spliceosome directs a set of proteins to bind to the mRNA near the position formerly occupied by the intron. Called the *exon junction complex (EJC),* these proteins mark the site of a successful splicing event and, as we shall see later in this chapter, influence the subsequent fate of the mRNA.

Other Properties of Pre-mRNA and Its Synthesis Help to Explain the Choice of Proper Splice Sites

As we have seen, intron sequences vary enormously in size, with some being in excess of 100,000 nucleotides. If splice-site selection were determined solely by the snRNPs acting on a preformed, protein-free RNA molecule, we would expect splicing mistakes—such as exon skipping and the use of "cryptic" splice sites—to be very common (**Figure 6–31**). The fidelity mechanisms built into the spliceosome, however, are supplemented by two additional strategies that increase the accuracy of splicing. The first is simply a consequence of the early stages of splicing occurring while the pre-mRNA molecules are being synthesized by RNA polymerase II. As transcription proceeds, the phosphorylated tail of RNA polymerase carries several components of the spliceosome (see Figure 6–23), and these components are transferred directly from the polymerase to the RNA as RNA is synthesized. This strategy helps the cell keep track of introns and exons: for example, the snRNPs that assemble at a 5′ splice site are initially presented with only a single 3′ splice site since the sites further downstream have not yet been synthesized. The coordination of transcription with splicing is especially important in preventing inappropriate exon skipping.

A strategy called "exon definition" is another way cells choose the appropriate splice sites. Exon size tends to be much more uniform than intron size, averaging about 150 nucleotide pairs across a wide variety of eucaryotic organisms (**Figure 6–32**). According to the exon definition idea, the splicing machinery initially seeks out the relatively homogenously sized exon sequences. As RNA synthesis proceeds, a group of additional components (most notably SR proteins, so-named because they contain a domain rich in serines and arginines) assemble on exon sequences and help to mark off each 3′ and 5′ splice site starting at the 5′ end of the RNA (**Figure 6–33**). These proteins, in turn, recruit U1 snRNA, which

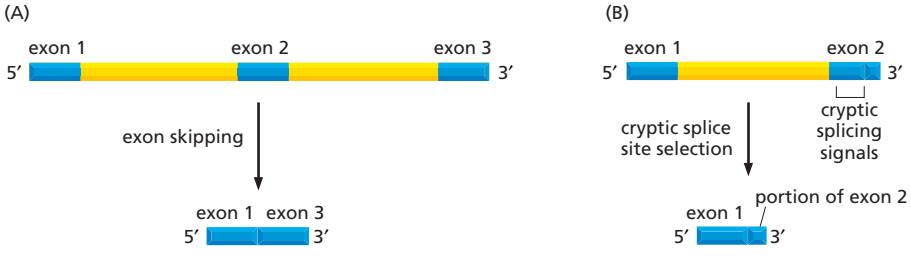


Figure 6–31 Two types of splicing errors. (A) Exon skipping. (B) Cryptic splice-site selection. Cryptic splicing signals are nucleotide sequences of RNA that closely resemble true splicing signals.

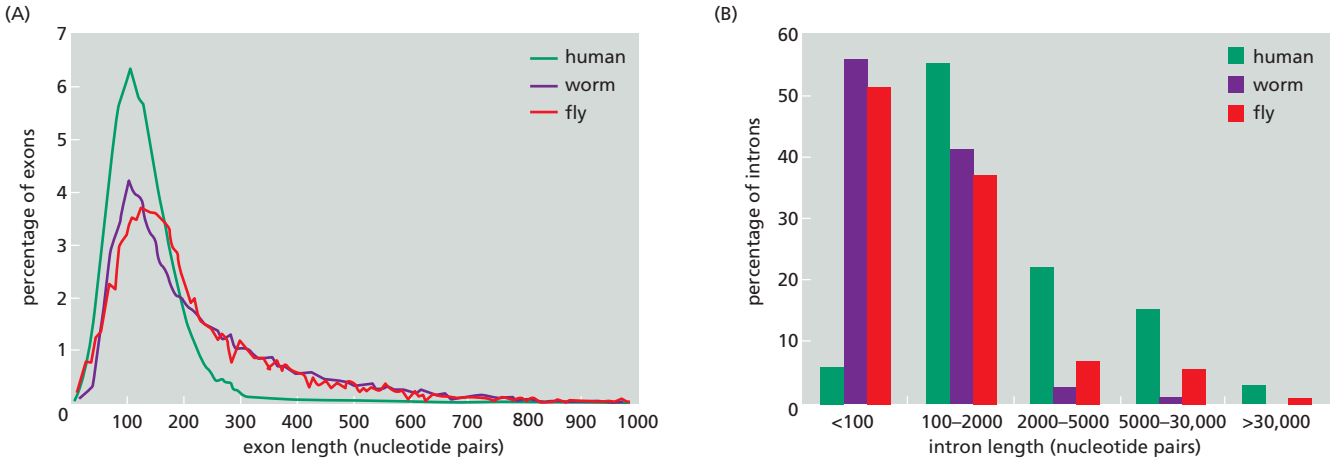


Figure 6–32 Variation in intron and exon lengths in the human, worm, and fly genomes. (A) Size distribution of exons. (B) Size distribution of introns. Note that exon length is much more uniform than intron length. (Adapted from International Human Genome Sequencing Consortium, *Nature* 409:860–921, 2001. With permission from Macmillan Publishers Ltd.)

marks the downstream exon boundary, and U2AF, which specifies the upstream one. By specifically marking the exons in this way and thereby taking advantage of the relatively uniform size of exons, the cell increases the accuracy with which it deposits the initial splicing components on the nascent RNA and thereby helps to avoid cryptic splice sites. How the SR proteins discriminate exon sequences from intron sequences is not understood in detail; however, it is known that some of the SR proteins bind preferentially to specific RNA sequences in exons, termed *splicing enhancers*. In principle, since any one of several different codons can be used to code for a given amino acid, there is freedom to adjust the exon nucleotide sequence so as to form a binding site for an SR protein, without necessarily affecting the amino acid sequence that the exon specifies.

Both the marking of exon and intron boundaries and the assembly of the spliceosome begin on an RNA molecule while it is still being elongated by RNA polymerase at its 3′ end. However, the actual chemistry of splicing can take place much later. This delay means that intron sequences are not necessarily removed from a pre-mRNA molecule in the order in which they occur along the RNA chain. It also means that, although spliceosome assembly is co-transcriptional, the splicing reactions sometimes occur posttranscriptionally—that is, after a complete pre-mRNA molecule has been made.

A Second Set of snRNPs Splice a Small Fraction of Intron Sequences in Animals and Plants

Simple eucaryotes such as yeasts have only one set of snRNPs that perform all pre-mRNA splicing. However, more complex eucaryotes such as flies, mammals, and plants have a second set of snRNPs that direct the splicing of a small fraction of their intron sequences. This minor form of spliceosome recognizes a different set of RNA sequences at the 5′ and 3′ splice junctions and at the branch point; it is called the *U12-type spliceosome* because of the involvement of the

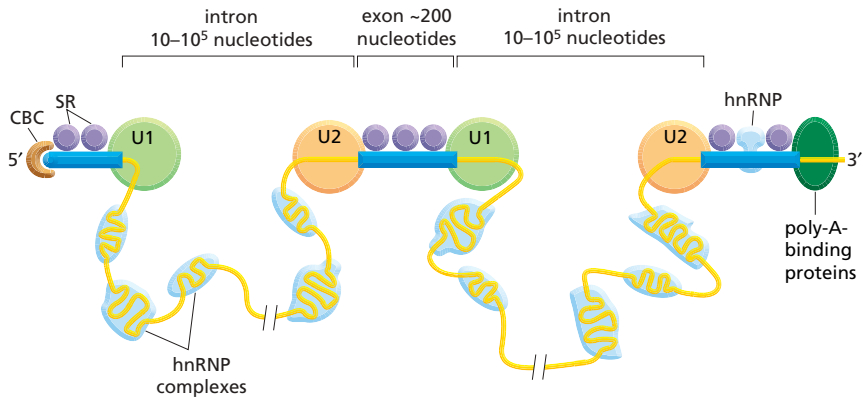


Figure 6–33 The exon definition idea. According to one proposal, SR proteins bind to each exon sequence in the pre-mRNA and thereby help to guide the snRNPs to the proper intron/exon boundaries. This demarcation of exons by the SR proteins occurs co-transcriptionally, beginning at the CBC (cap-binding complex) at the 5′ end. As indicated, the intron sequences in the pre-mRNA, which can be extremely long, are packaged into hnRNP (heterogeneous nuclear ribonucleoprotein) complexes that compact them into more manageable structures and perhaps mask cryptic splice sites. It has been proposed that hnRNP proteins may preferentially associate with intron sequences and that this preference may also help the spliceosome distinguish introns from exons. However, as shown, at least some hnRNP proteins also bind to exon sequences. (Adapted from R. Reed, *Curr. Opin. Cell Biol.* 12:340–345, 2000. With permission from Elsevier.)

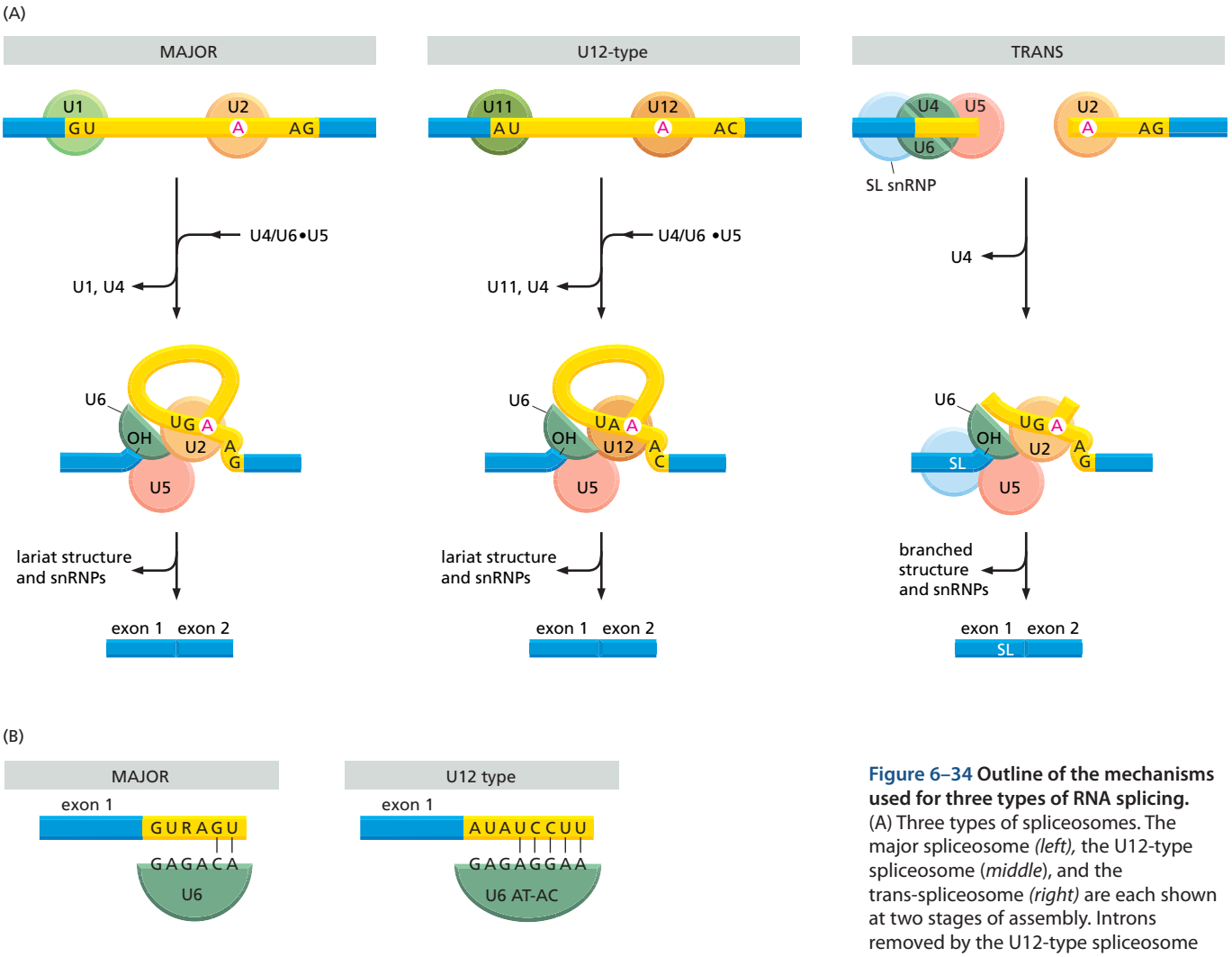


Figure 6–34 Outline of the mechanisms used for three types of RNA splicing. (A) Three types of spliceosomes. The major spliceosome *(left),* the U12-type spliceosome (*middle*), and the trans-spliceosome *(right)* are each shown at two stages of assembly. Introns removed by the U12-type spliceosome have a different set of consensus nucleotide sequences from those removed by the major spliceosome. In humans, it is estimated that 0.1% of introns are removed by the U12-type spliceosome. In trans-splicing, which does not occur in humans, the SL snRNP is consumed in the reaction because a portion of the SL snRNA becomes the first exon of the mature mRNA. (B) The major U6 snRNP and the U6 snRNP specific to the U12-type spliceosome both recognize the 5′ splice junction, but they do so through a different set of base-pair interactions. The sequences shown are from humans. (Adapted from Y.T. Yu et al., The RNA World, pp. 487–524. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press, 1999.)

U12 SnRNP (**Figure 6–34**A). Despite recognizing different nucleotide sequences, the snRNPs in this spliceosome make the same types of RNA–RNA interactions with the pre-mRNA and with each other as do the major snRNPs (Figure 6–34B). Although, as we have seen, components of the major spliceosomes travel with RNA polymerase II as it transcribes genes, this may not be the case for the U12 spliceosome. It is possible that U12-mediated splicing is thereby delayed, and this presents the cell with a way to co-regulate splicing of the several hundred genes whose expression requires this spliceosome. A number of mammalian mRNAs contain a mixture of introns, some removed by the major spliceosome and others by the minor spliceosome, and it has been proposed that this arrangement permits particularly complex patterns of alternative splicing to occur.

A few eucaryotic organisms exhibit a particular variation on splicing, called **trans-splicing**. These organisms include the single-celled trypanosomes—protozoans that cause African sleeping sickness in humans—and the model multicellular organism, the nematode worm. In trans-splicing, exons from two separate RNA transcripts are spliced together to form a mature mRNA molecule (see Figure 6–34A). Trypanosomes produce all of their mRNAs in this way, whereas trans-splicing accounts for only about 1% of nematode mRNAs. In both cases, a single exon is spliced onto the 5′ end of many different RNA transcripts produced by the cell; in this way, all of the products of trans-splicing have the same 5′ exon and different 3′ exons. Many of the same snRNPs that function in conventional splicing are used in this reaction, although trans-splicing uses a unique snRNP (called the SL RNP) that brings in the common exon (see Figure 6–34).

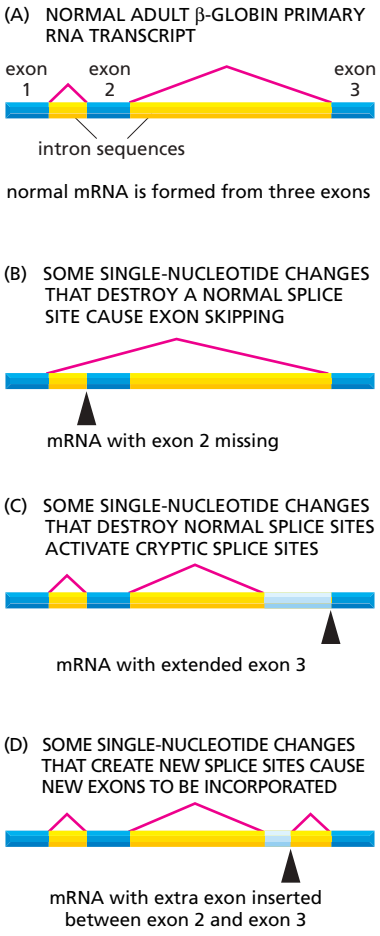


Figure 6–35 Abnormal processing of the β-globin primary RNA transcript in humans with the disease β thalassemia. In the examples shown, the disease is caused by splice-site mutations (*black arrowheads*) found in the genomes of affected patients. The *dark blue boxes* represent the three normal exon sequences; the *red lines* indicate the 5′ and 3′ splice sites. The *light blue boxes* depict new nucleotide sequences included in the final mRNA molecule as a result of the mutation. Note that when a mutation leaves a normal splice site without a partner, an exon is skipped or one or more abnormal cryptic splice sites nearby is used as the partner site, as in (C) and (D). (Adapted in part from S.H. Orkin, in The Molecular Basis of Blood Diseases [G. Stamatoyannopoulos et al., eds.], pp. 106–126. Philadelphia: Saunders, 1987.)

We do not know why even a few organisms use trans-splicing; however, it is thought that the common 5′ exon may aid in the translation of the mRNA. Thus, the mRNAs produced by trans-splicing in nematodes seem to be translated with especially high efficiency.

RNA Splicing Shows Remarkable Plasticity

We have seen that the choice of splice sites depends on such features of the pre-mRNA transcript as the affinity of the three signals on the RNA (the 5′ and 3′ splice junctions and the branch point) for the splicing machinery, the co-transcriptional assembly of the spliceosome, and the "bookkeeping" that underlies exon definition. We do not know how accurate splicing normally is because, as we see later, there are several quality control systems that rapidly destroy mRNAs whose splicing goes awry. However, we do know that, compared with other steps in gene expression, splicing is unusually flexible. For example, a mutation in a nucleotide sequence critical for splicing of a particular intron does not necessarily prevent splicing of that intron altogether. Instead, the mutation typically creates a new pattern of splicing (**Figure 6–35**). Most commonly, an exon is simply skipped (Figure 6–35B). In other cases, the mutation causes a cryptic splice junction to be efficiently used (Figure 6–35C). Apparently, the splicing machinery has evolved to pick out the best possible pattern of splice junctions, and if the optimal one is damaged by mutation, it will seek out the next best pattern, and so on. This flexibility in the process of RNA splicing suggests that changes in splicing patterns caused by random mutations have been an important pathway in the evolution of genes and organisms.

The plasticity of RNA splicing also means that the cell can regulate the pattern of RNA splicing. Earlier in this section we saw that alternative splicing can give rise to different proteins from the same gene. Some examples of alternative splicing are constitutive; that is, the alternatively spliced mRNAs are produced continuously by cells of an organism. However, in many cases, the cell regulates the splicing patterns so that different forms of the protein are produced at different times and in different tissues (see Figure 6–27). In Chapter 7 we return to this issue to discuss some specific examples of regulated RNA splicing.

Spliceosome-Catalyzed RNA Splicing Probably Evolved from Self-splicing Mechanisms

When the spliceosome was first discovered, it puzzled molecular biologists. Why do RNA molecules instead of proteins perform important roles in splice site recognition and in the chemistry of splicing? Why is a lariat intermediate used rather than the apparently simpler alternative of bringing the 5′ and 3′ splice sites together in a single step, followed by their direct cleavage and rejoining? The answers to these questions reflect the way in which the spliceosome is believed to have evolved.

As discussed briefly in Chapter 1 (and in more detail in the final section of this chapter), it is likely that early cells used RNA molecules rather than proteins as their major catalysts and that they stored their genetic information in RNA rather than in DNA sequences. RNA-catalyzed splicing reactions presumably had important roles in these early cells. As evidence, some *self-splicing RNA* introns (that is, intron sequences in RNA whose splicing out can occur in the absence of proteins or any other RNA molecules) remain today—for example, in the nuclear rRNA genes of the ciliate *Tetrahymena*, in a few bacteriophage T4 genes, and in some mitochondrial and chloroplast genes.

A self-splicing intron sequence can be identified in a test tube by incubating a pure RNA molecule that contains the intron sequence and observing the splicing reaction. Two major classes of self-splicing intron sequences can be distinguished in this way. *Group I intron sequences* begin the splicing reaction by binding a G nucleotide to the intron sequence; this G is thereby activated to form the attacking group that will break the first of the phosphodiester bonds cleaved during splicing (the bond at the 5′ splice site). In *group II intron sequences,* an especially reactive A residue in the intron sequence is the attacking group, and a lariat intermediate is generated. Otherwise the reaction pathways for the two types of self-splicing intron sequences are the same. Both are presumed to represent vestiges of very ancient mechanisms (**Figure 6–36**).

For both types of self-splicing reactions, the nucleotide sequence of the intron is critical; the intron RNA folds into a specific three-dimensional structure, which brings the 5′ and 3′ splice junctions together and provides precisely positioned reactive groups to perform the chemistry (see Figure 6–6C). Because the chemistries of their splicing reactions are so similar, it has been proposed that the pre-mRNA splicing mechanism of the spliceosome evolved from group

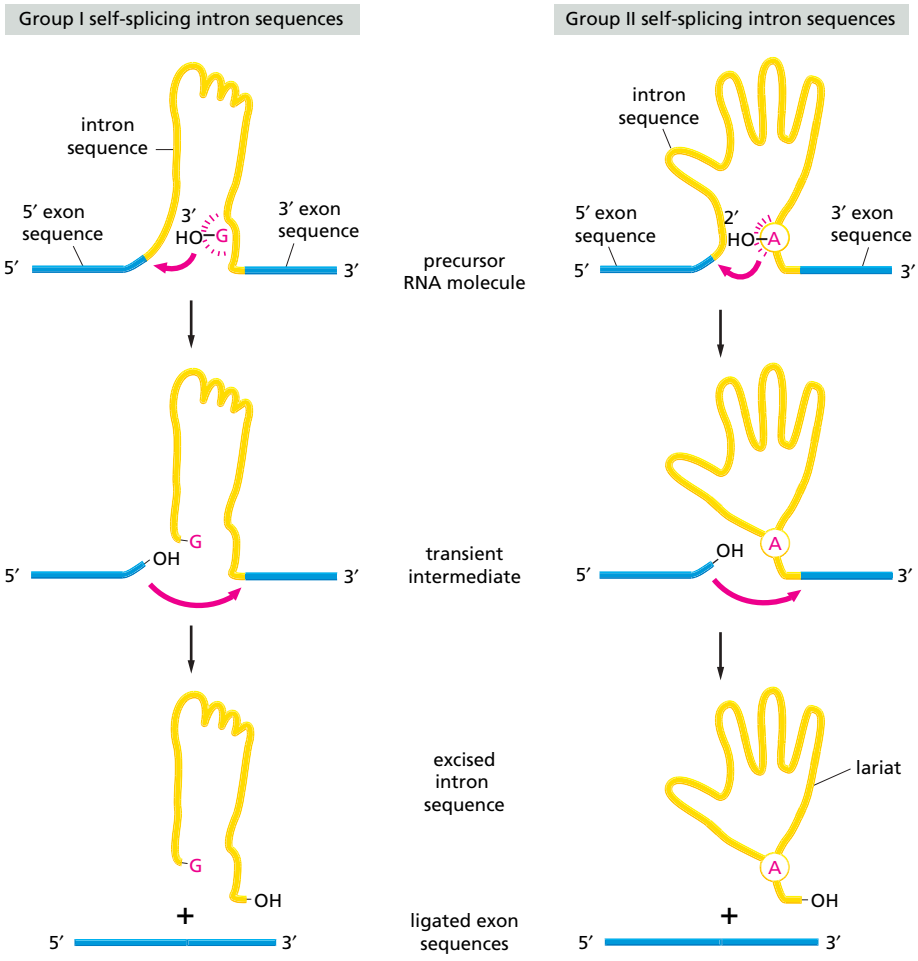


Figure 6–36 The two known classes of self-splicing intron sequences. The figure emphasizes the similarities between the two mechanisms. Both are normally aided by proteins in the cell that speed up the reaction, but the catalysis is nevertheless mediated by the RNA in the intron sequence. The group I intron sequences bind a free G nucleotide to a specific site on the RNA to initiate splicing, while the group II intron sequences use an especially reactive A nucleotide in the intron sequence itself for the same purpose. Both types of self-splicing reactions require the intron to be folded into a highly specific three-dimensional structure that provides the catalytic activity for the reaction (see Figure 6–6). The mechanism used by group II intron sequences releases the intron as a lariat structure and closely resembles the pathway of pre-mRNA splicing catalyzed by the spliceosome (compare with Figure 6–29). The spliceosome performs most RNA splicing in eucaryotic cells, and self-splicing RNAs represent unusual cases. (Adapted from T.R. Cech, *Cell* 44:207–210, 1986. With permission from Elsevier.)

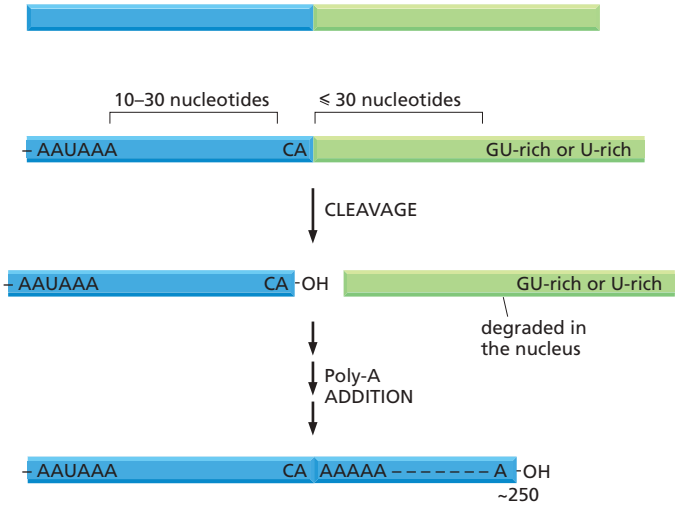


Figure 6–37 Consensus nucleotide sequences that direct cleavage and polyadenylation to form the 3′ end of a eucaryotic mRNA. These sequences are encoded in the genome; specific proteins recognize them after they are transcribed into RNA. The hexamer AAUAAA is bound by CPSF, the GU-rich element beyond the cleavage site is bound by CstF (see Figure 6–38), and the CA sequence is bound by a third factor required for the cleavage step. Like other consensus nucleotide sequences discussed in this chapter (see Figure 6–12), the sequences shown in the figure represent a variety of individual cleavage and polyadenylation signals.

II self-splicing. According to this idea, when the spliceosomal snRNPs took over the structural and chemical roles of the group II introns, the strict sequence constraints on intron sequences would have disappeared, thereby permitting a vast expansion in the number of different RNAs that could be spliced.

RNA-Processing Enzymes Generate the 3′ End of Eucaryotic mRNAs

As previously explained, the 5′ end of the pre-mRNA produced by RNA polymerase II is capped almost as soon as it emerges from the RNA polymerase. Then, as the polymerase continues its movement along a gene, the spliceosome assembles on the RNA and delineates the intron and exon boundaries. The long C-terminal tail of the RNA polymerase coordinates these processes by transferring capping and splicing components directly to the RNA as it emerges from the enzyme. We see in this section that, as RNA polymerase II reaches the end of a gene, a similar mechanism ensures that the 3′ end of the pre-mRNA is appropriately processed.

As might be expected, the position of the 3′ end of each mRNA molecule is ultimately specified by a signal encoded in the genome (**Figure 6–37**). These signals are transcribed into RNA as the RNA polymerase II moves through them, and they are then recognized (as RNA) by a series of RNA-binding proteins and RNA-processing enzymes (**Figure 6–38**). Two multisubunit proteins, called CstF (cleavage stimulation factor) and CPSF (cleavage and polyadenylation specificity factor), are of special importance. Both of these proteins travel with the RNA polymerase tail and are transferred to the 3′-end processing sequence on an RNA molecule as it emerges from the RNA polymerase.

Once CstF and CPSF bind to specific nucleotide sequences on the emerging RNA molecule, additional proteins assemble with them to create the 3′ end of the mRNA. First, the RNA is cleaved (see Figure 6–38). Next an enzyme called poly-A polymerase (PAP) adds, one at a time, approximately 200 A nucleotides to the 3′ end produced by the cleavage. The nucleotide precursor for these additions is ATP, and the same type of 5′-to-3′ bonds are formed as in conventional RNA synthesis (see Figure 6–4). Unlike the usual RNA polymerases, poly-A polymerase does not require a template; hence the poly-A tail of eucaryotic mRNAs

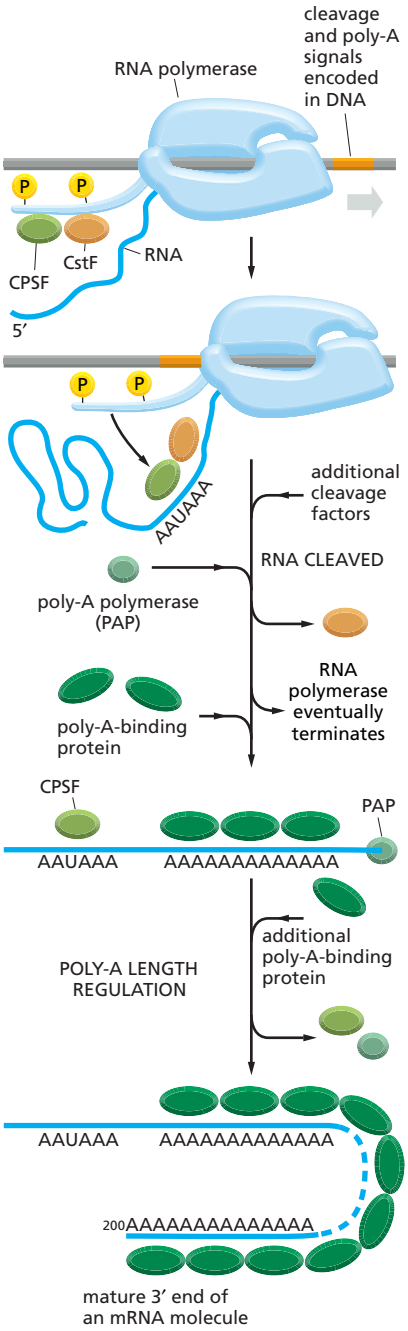


Figure 6–38 Some of the major steps in generating the 3′ end of a eucaryotic mRNA. This process is much more complicated than the analogous process in bacteria, where the RNA polymerase simply stops at a termination signal and releases both the 3′ end of its transcript and the DNA template (see Figure 6–11).

is not directly encoded in the genome. As the poly-A tail is synthesized, proteins called poly-A-binding proteins assemble onto it and, by a poorly understood mechanism, determine the final length of the tail. Some poly-A-binding proteins remain bound to the poly-A tail as the mRNA travels from the nucleus to the cytosol and they help to direct the synthesis of a protein on the ribosome, as we see later in this chapter.

After the 3′ end of a eucaryotic pre-mRNA molecule has been cleaved, the RNA polymerase II continues to transcribe, in some cases for hundreds of nucleotides. But the polymerase soon releases its grip on the template and transcription terminates. After 3′-end cleavage has occurred, the newly synthesized RNA that emerges from the polymerases lacks a 5′ cap; this unprotected RNA is rapidly degraded by a 5′ → 3′ exonuclease, which is carried along on the polymerase tail. Apparently, it is this RNA degradation that eventually causes the RNA polymerase to dissociate from the DNA.

Mature Eucaryotic mRNAs Are Selectively Exported from the Nucleus

We have seen how eucaryotic pre-mRNA synthesis and processing take place in an orderly fashion within the cell nucleus. However, these events create a special problem for eucaryotic cells, especially those of complex organisms where the introns are vastly longer than the exons. Of the pre-mRNA that is synthesized, only a small fraction—the mature mRNA—is of further use to the cell. The rest—excised introns, broken RNAs, and aberrantly processed pre-mRNAs—is not only useless but potentially dangerous. How, then, does the cell distinguish between the relatively rare mature mRNA molecules it wishes to keep and the overwhelming amount of debris from RNA processing?

The answer is that, as an RNA molecule is processed, it loses certain proteins and acquires others, thereby signifying the successful completion of each of the different steps. For example, we have seen that acquisition of the cap-binding complexes, the exon junction complexes, and the poly-A-binding proteins mark the completion of capping, splicing, and poly-A addition, respectively. A properly completed mRNA molecule is also distinguished by the proteins it lacks. For example, the presence of a snRNP would signify incomplete or aberrant splicing. Only when the proteins present on an mRNA molecule collectively signify that processing was successfully completed is the mRNA exported from the nucleus into the cytosol, where it can be translated into protein. Improperly processed mRNAs, and other RNA debris are retained in the nucleus, where they are eventually degraded by the nuclear **exosome**, a large protein complex whose interior is rich in 3′-to-5′ RNA exonucleases. Eucaryotic cells thus export only useful RNA molecules to the cytoplasm, while debris is disposed of in the nucleus.

Of all the proteins that assemble on pre-mRNA molecules as they emerge from transcribing RNA polymerases, the most abundant are the hnRNPs (heterogeneous nuclear ribonuclear proteins) (see Figure 6–33). Some of these proteins (there are approximately 30 of them in humans) unwind the hairpin helices from the RNA so that splicing and other signals on the RNA can be read more easily. Others preferentially package the RNA contained in the very long intron sequences typically found in genes of complex organisms. They may therefore play an important role in distinguishing mature mRNA from the debris left over from RNA processing.

Successfully processed mRNAs are guided through the **nuclear pore complexes** (NPCs)—aqueous channels in the nuclear membrane that directly connect the nucleoplasm and cytosol (**Figure 6–39**). Small molecules (less than 50,000 daltons) can diffuse freely through these channels. However, most of the macromolecules in cells, including mRNAs complexed with proteins, are far too large to pass through the channels without a special process. The cell uses energy to actively transport such macromolecules in both directions through the nuclear pore complexes.

As explained in detail in Chapter 12, macromolecules are moved through nuclear pore complexes by *nuclear transport receptors,* which, depending on the

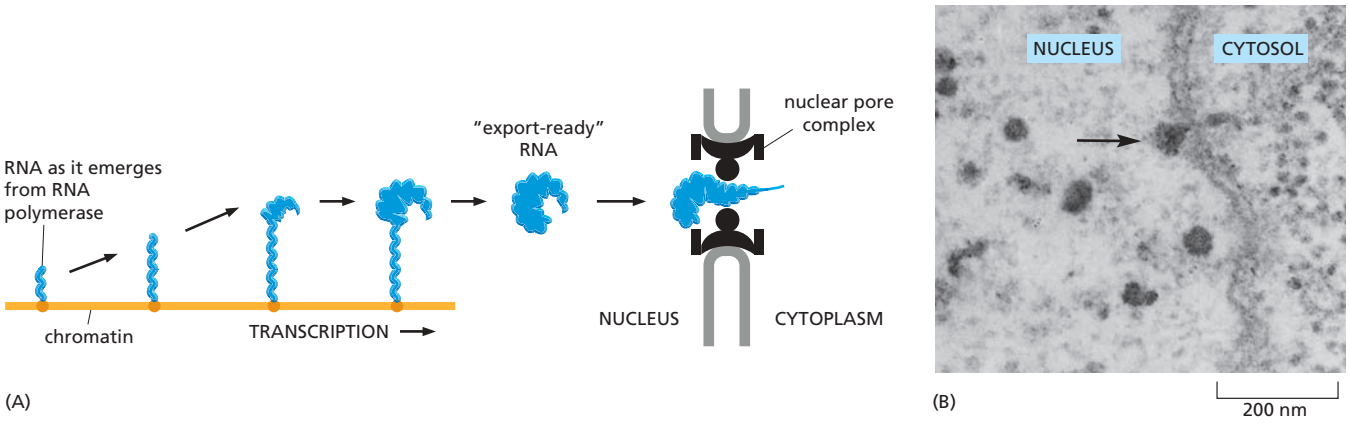


Figure 6–39 Transport of a large mRNA molecule through the nuclear pore complex. (A) The maturation of an mRNA molecule as it is synthesized by RNA polymerase and packaged by a variety of nuclear proteins. This drawing of an unusually abundant RNA, called the Balbiani Ring mRNA, is based on EM micrographs such as that shown in (B). Balbiani Rings are found in the cells of certain insects. (A, adapted from B. Daneholt, *Cell* 88:585–588, 1997. With permission from Elsevier; B, from B.J. Stevens and H. Swift, *J. Cell Biol*. 31:55–77, 1966. With permission from The Rockefeller University Press.)

identity of the macromolecule, escort it from the nucleus to the cytoplasm or vice versa. For mRNA export to occur, a specific nuclear transport receptor must be loaded onto the mRNA, a step that, at least in some organisms, takes place in concert with 3′ cleavage and polyadenylation. Once it helps to move an RNA molecule through the nuclear pore complex, the transport receptor dissociates from the mRNA, re-enters the nucleus, and exports a new mRNA molecule (**Figure 6–40**).

The export of mRNA–protein complexes from the nucleus can be observed with the electron microscope for the unusually abundant mRNA of the insect Balbiani Ring genes. As these genes are transcribed, the newly formed RNA is seen to be packaged by proteins, including hnRNPs, SR proteins, and components of the spliceosome. This protein–RNA complex undergoes a series of structural transitions, probably reflecting RNA processing events, culminating in a curved fiber (see Figure 6–39). This curved fiber moves through the nucleoplasm and enters the nuclear pore complex (with its 5′ cap proceeding first), and it then undergoes another series of structural transitions as it moves through the

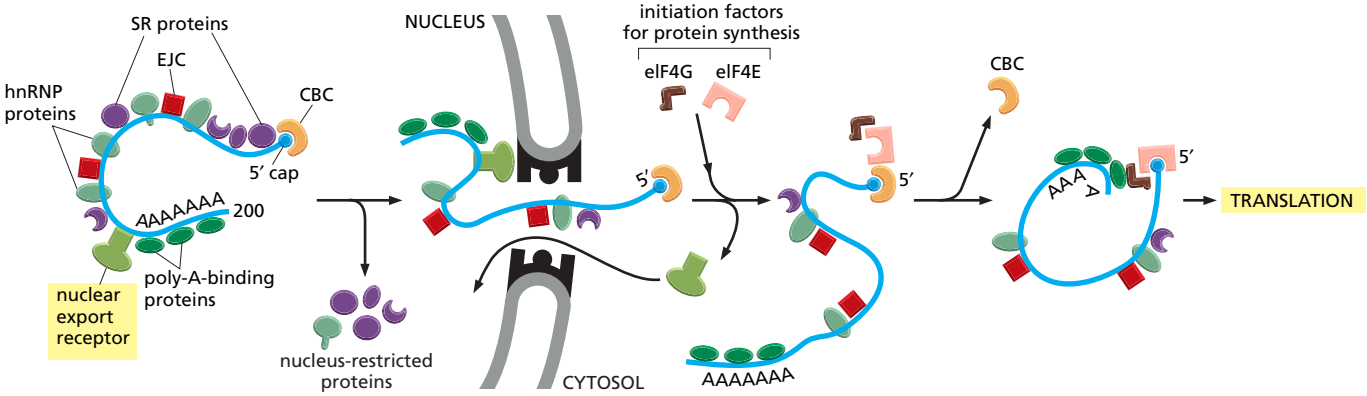


Figure 6–40 Schematic illustration of an "export-ready" mRNA molecule and its transport through the nuclear pore. As indicated, some proteins travel with the mRNA as it moves through the pore, whereas others remain in the nucleus. The nuclear export receptor for mRNAs is a complex of proteins that is deposited when the mRNA has been correctly spliced and polyadenylated. When the mRNA is exported to the cytosol, the export receptor dissociates from the mRNA and is re-imported into the nucleus, where it can be used again. Just after it leaves the nucleus, and before it loses the cap-binding complex (CBC) the mRNA is subjected to a final check, called *nonsense-mediated decay*, which is described later in the chapter. Once it passes this test the mRNA continues to shed previously bound proteins and acquire new ones before it is efficiently translated into protein. EJC, exon junction complex.

pore. These and other observations reveal that the pre-mRNA–protein and mRNA–protein complexes are dynamic structures that gain and lose numerous specific proteins during RNA synthesis, processing, and export (see Figure 6–40).

As we have seen, some of these proteins mark the different stages of mRNA maturation; other proteins deposited on the mRNA while it is still in the nucleus can affect the fate of the RNA after it is transported to the cytosol. Thus, the stability of an mRNA in the cytosol, the efficiency with which it is translated into protein, and its ultimate destination in the cytosol can all be determined by proteins acquired in the nucleus that remain bound to the RNA after it leaves the nucleus. We will discuss these issues in Chapter 7 when we turn to the post-transcriptional control of gene expression.

We have seen that RNA synthesis and processing are closely coupled in the cell, and it might be expected that export from the nucleus is somehow integrated with these two processes. Although the Balbiani Ring RNAs can be seen to move through the nucleoplasm and out through the nuclear pores, other mRNAs appear to be synthesized and processed in close proximity to nuclear pore complexes. In these cases, which may represent the majority of eucaryotic genes, mRNA synthesis, processing, and transport all appear to be tightly coupled; the mRNA can thus be viewed as emerging from the nuclear pore as a newly manufactured car might emerge from an assembly line. Later in this chapter, we will see that the cell performs an additional quality-control check on each mRNA before it is allowed to be efficiently translated into protein.

Before discussing what happens to mRNAs after they leave the nucleus, we briefly consider how the synthesis and processing of noncoding RNA molecules occurs. Although there are many other examples, our discussion focuses on the rRNAs that are critically important for the translation of mRNAs into protein.

Many Noncoding RNAs Are Also Synthesized and Processed in the Nucleus

A few percent of the dry weight of a mammalian cell is RNA; of that, only about 3–5% is mRNA. A fraction of the remainder represents intron sequences before they have been degraded, but the bulk of the RNA in cells performs structural and catalytic functions (see Table 6–1, p. 336). The most abundant RNAs in cells are the ribosomal RNAs (rRNAs), constituting approximately 80% of the RNA in rapidly dividing cells. As discussed later in this chapter, these RNAs form the core of the ribosome. Unlike bacteria—in which a single RNA polymerase synthesizes all RNAs in the cell—eucaryotes have a separate, specialized polymerase, RNA polymerase I, that is dedicated to producing rRNAs. RNA polymerase I is similar structurally to the RNA polymerase II discussed previously; however, the absence of a C-terminal tail in polymerase I helps to explain why its transcripts are neither capped nor polyadenylated. As mentioned earlier, this difference helps the cell distinguish between noncoding RNAs and mRNAs.

Because multiple rounds of translation of each mRNA molecule can provide an enormous amplification in the production of protein molecules, many of the proteins that are very abundant in a cell can be synthesized from genes that are present in a single copy per haploid genome. In contrast, the RNA components of the ribosome are final gene products, and a growing mammalian cell must synthesize approximately 10 million copies of each type of ribosomal RNA in each cell generation to construct its 10 million ribosomes. The cell can produce adequate quantities of ribosomal RNAs only because it contains multiple copies of the **rRNA genes** that code for ribosomal RNAs (**rRNAs**). Even *E. coli* needs seven copies of its rRNA genes to meet the cell's need for ribosomes. Human cells contain about 200 rRNA gene copies per haploid genome, spread out in small clusters on five different chromosomes (see Figure 4–11), while cells of the frog *Xenopus* contain about 600 rRNA gene copies per haploid genome in a single cluster on one chromosome (**Figure 6–41**).

There are four types of eucaryotic rRNAs, each present in one copy per ribosome. Three of the four rRNAs (18S, 5.8S, and 28S) are made by chemically modifying and cleaving a single large precursor rRNA (**Figure 6–42**); the fourth (5S

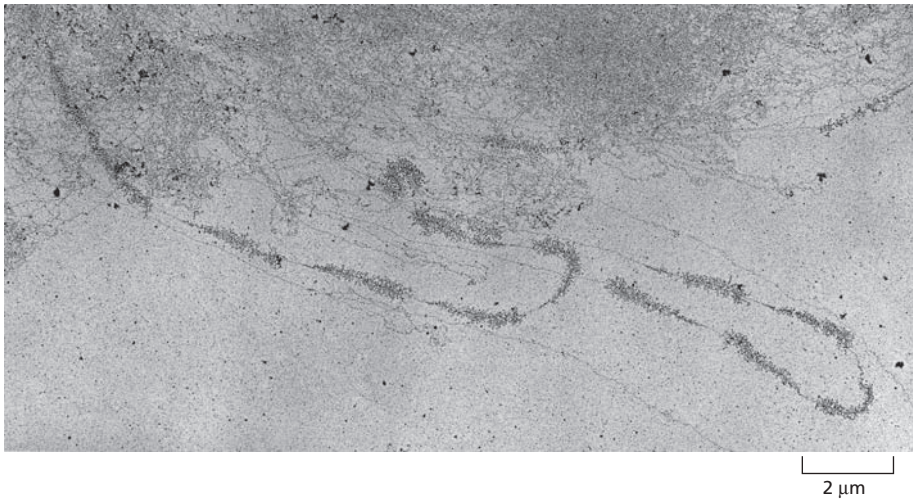


Figure 6–41 Transcription from tandemly arranged rRNA genes, as seen in the electron microscope. The pattern of alternating transcribed gene and nontranscribed spacer is readily seen. A higher-magnification view of rRNA genes is shown in Figure 6–9. (From V.E. Foe, *Cold Spring Harbor Symp. Quant. Biol*. 42:723–740, 1978. With permission from Cold Spring Harbor Laboratory Press.)

RNA) is synthesized from a separate cluster of genes by a different polymerase, RNA polymerase III, and does not require chemical modification.

Extensive chemical modifications occur in the 13,000-nucleotide-long precursor rRNA before the rRNAs are cleaved out of it and assembled into ribosomes. These include about 100 methylations of the 2′-OH positions on nucleotide sugars and 100 isomerizations of uridine nucleotides to pseudouridine (**Figure 6–43**A). The functions of these modifications are not understood in detail, but many probably aid in the folding and assembly of the final rRNAs and some may subtly alter the function of ribosomes. Each modification is made at a specific position in the precursor rRNA. These positions are specified by about 150 "guide RNAs," which position themselves through base-pairing to the precursor rRNA and thereby bring an RNA-modifying enzyme to the appropriate position (Figure 6–43B). Other guide RNAs promote cleavage of the precursor rRNAs into the mature rRNAs, probably by causing conformational changes in the precursor rRNA that expose these sites to nucleases. All of these guide RNAs are members of a large class of RNAs called **small nucleolar RNAs** (or **snoRNAs**), so named because these RNAs perform their functions in a subcompartment of the nucleus called the nucleolus. Many snoRNAs are encoded in the introns of other genes, especially those encoding ribosomal proteins. They are therefore synthesized by RNA polymerase II and processed from excised intron sequences.

Recently several snoRNA-like RNAs have been identified that are synthesized only in cells of the brain. These are believed to direct the modification of mRNAs, instead of rRNAs, and are likely to represent a new, but poorly understood, type of gene regulatory mechanism.

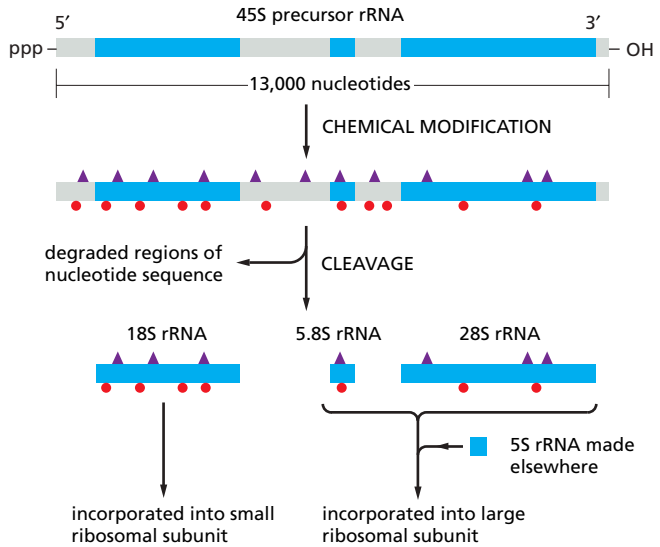


Figure 6–42 The chemical modification and nucleolytic processing of a eucaryotic 45S precursor rRNA molecule into three separate ribosomal RNAs. Two types of chemical modifications (color-coded as indicated in Figure 6–43) are made to the precursor rRNA before it is cleaved. Nearly half of the nucleotide sequences in this precursor rRNA are discarded and degraded in the nucleus. The rRNAs are named according to their "S" values, which refer to their rate of sedimentation in an ultracentrifuge. The larger the S value, the larger the rRNA.

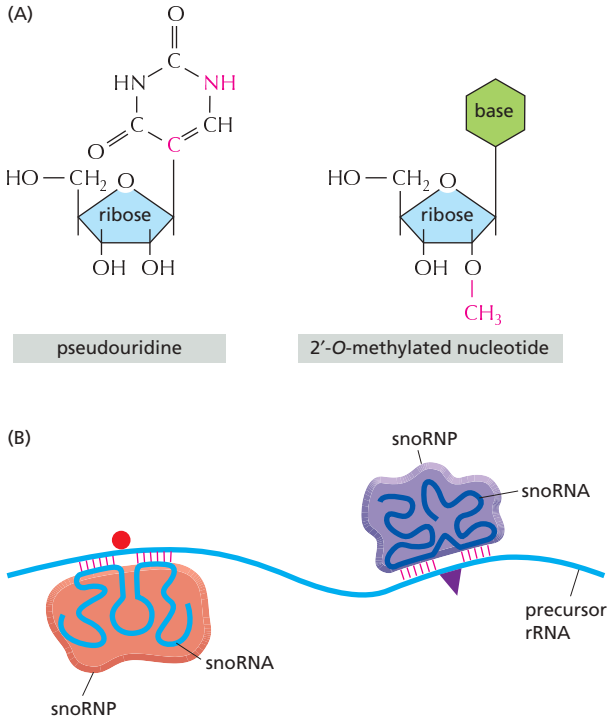
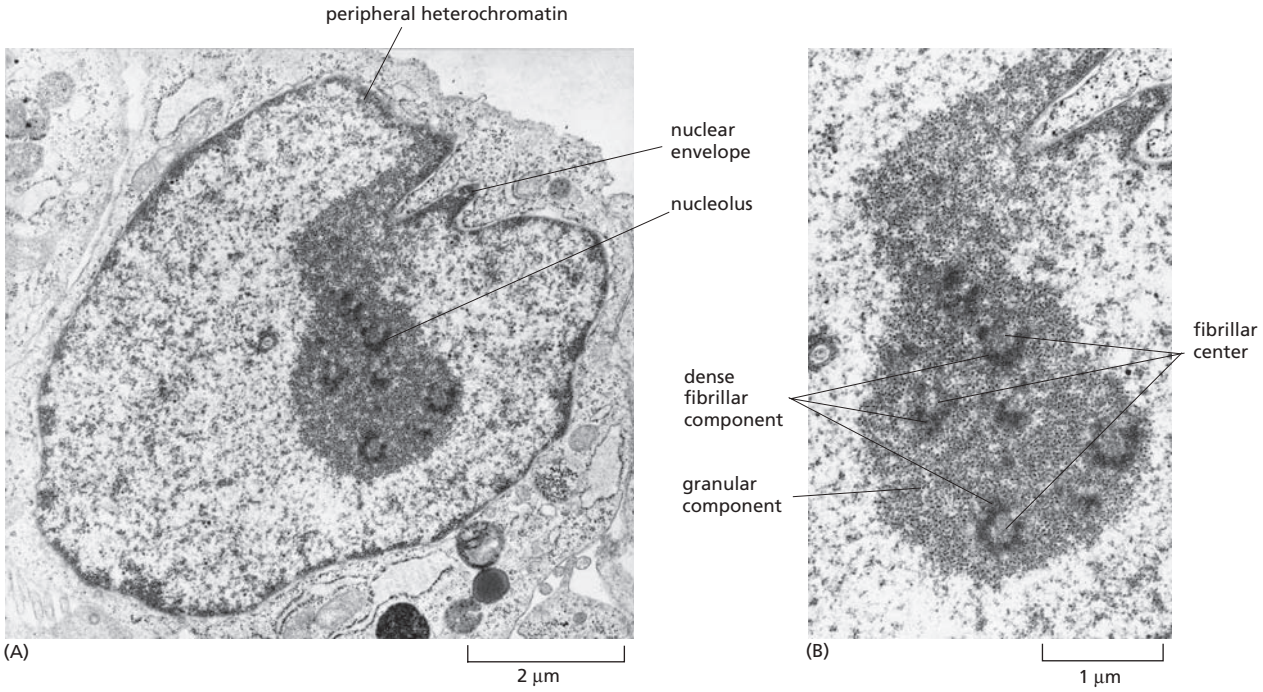


Figure 6–43 Modifications of the precursor rRNA by guide RNAs. (A) Two prominent covalent modifications occur after rRNA synthesis; the differences from the initially incorporated nucleotide are indicated by *red* atoms. Pseudouridine is an isomer of uridine; the base has been "rotated" relative to the sugar. (B) As indicated, snoRNAs determine the sites of modification by base-pairing to complementary sequences on the precursor rRNA. The snoRNAs are bound to proteins, and the complexes are called snoRNPs. snoRNPs contain both the guide sequences and the enzymes that modify the rRNA.

The Nucleolus Is a Ribosome-Producing Factory

The nucleolus is the most obvious structure seen in the nucleus of a eucaryotic cell when viewed in the light microscope. Consequently, it was so closely scrutinized by early cytologists that an 1898 review could list some 700 references. We now know that the nucleolus is the site for the processing of rRNAs and their assembly into ribosome subunits. Unlike many of the major organelles in the cell, the nucleolus is not bound by a membrane (**Figure 6–44**); instead, it is a large aggregate of macromolecules, including the rRNA genes themselves, precursor rRNAs, mature rRNAs, rRNA-processing enzymes, snoRNPs, ribosomal proteins and partly assembled ribosomes. The close association of all these components presumably allows the assembly of ribosomes to occur rapidly and smoothly.

Figure 6–44 Electron micrograph of a thin section of a nucleolus in a human fibroblast, showing its three distinct zones. (A) View of entire nucleus. (B) High-power view of the nucleolus. It is believed that transcription of the rRNA genes takes place between the fibrillar center and the dense fibrillar component and that processing of the rRNAs and their assembly into the two subunits of the ribosome proceeds outward from the dense fibrillar component to the surrounding granular components. (Courtesy of E.G. Jordan and J. McGovern.)



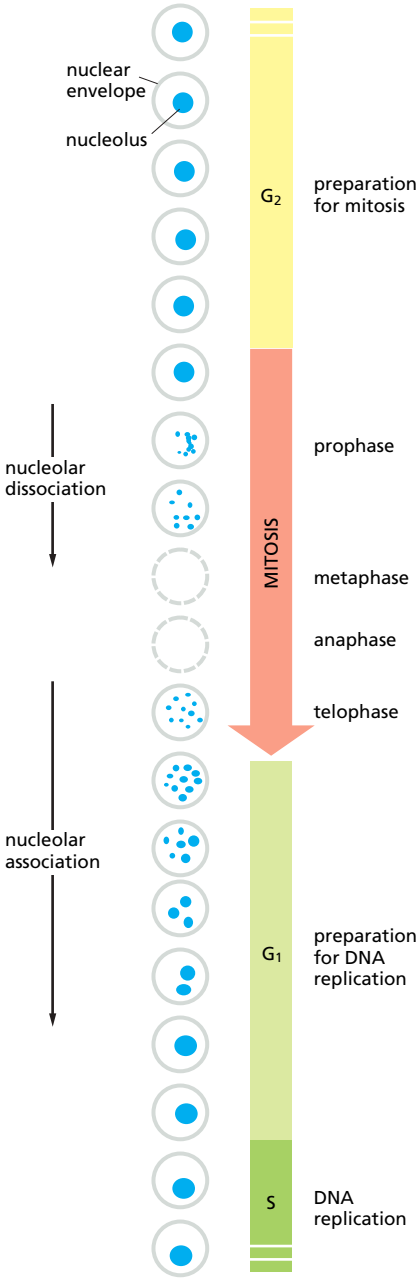


Figure 6–45 Changes in the appearance of the nucleolus in a human cell during the cell cycle. Only the cell nucleus is represented in this diagram. In most eucaryotic cells the nuclear envelope breaks down during mitosis, as indicated by the dashed circles.

Various types of RNA molecules play a central part in the chemistry and structure of the nucleolus, suggesting that it may have evolved from an ancient structure present in cells dominated by RNA catalysis. In present-day cells, the rRNA genes also have an important role in forming the nucleolus. In a diploid human cell, the rRNA genes are distributed into 10 clusters, located near the tips of five different chromosome pairs (see Figure 4–11). During interphase these 10 chromosomes contribute DNA loops (containing the rRNA genes) to the nucleolus; in M-phase, when the chromosomes condense, the nucleolus disappears. Finally, in the telophase part of mitosis, as chromosomes return to their semi-dispersed state, the tips of the 10 chromosomes coalesce and the nucleolus reforms (**Figure 6–45** and **Figure 6–46**). The transcription of the rRNA genes by RNA polymerase I is necessary for this process. As might be expected, the size of the nucleolus reflects the number of ribosomes that the cell is producing. Its size therefore varies greatly in different cells and can change in a single cell, occupying 25% of the total nuclear volume in cells that are making unusually large amounts of protein.

Ribosome assembly is a complex process, the most important features of which are outlined in **Figure 6–47**. In addition to its important role in ribosome biogenesis, the nucleolus is also the site where other RNAs are produced and other RNA–protein complexes are assembled. For example, the U6 snRNP, which functions in pre-mRNA splicing (see Figure 6–29), is composed of one RNA molecule and at least seven proteins. The U6 snRNA is chemically modified by snoRNAs in the nucleolus before its final assembly there into the U6 snRNP. Other important RNA–protein complexes, including telomerase (encountered in Chapter 5) and the signal recognition particle (which we discuss in Chapter 12), are also believed to be assembled at the nucleolus. Finally, the tRNAs (transfer RNAs) that carry the amino acids for protein synthesis are processed there as well; like the rRNA genes, those encoding tRNAs are clustered in the nucleolus. Thus, the nucleolus can be thought of as a large factory at which many different noncoding RNAs are transcribed, processed, and assembled with proteins to form a large variety of ribonucleoprotein complexes.

The Nucleus Contains a Variety of Subnuclear Structures

Although the nucleolus is the most prominent structure in the nucleus, several other nuclear bodies have been observed and studied (**Figure 6–48**). These include Cajal bodies (named for the scientist who first described them in 1906), GEMS (Gemini of Cajal bodies), and interchromatin granule clusters (also called

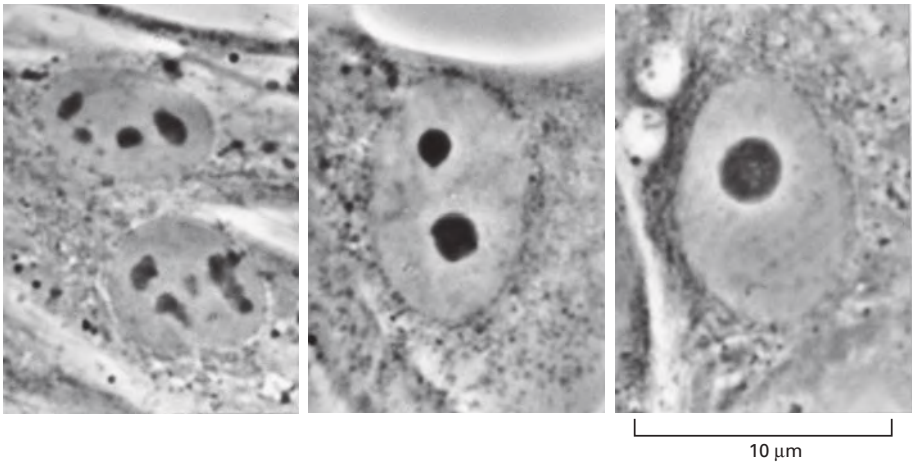


Figure 6–46 Nucleolar fusion. These light micrographs of human fibroblasts grown in culture show various stages of nucleolar fusion. After mitosis, each of the 10 human chromosomes that carry a cluster of rRNA genes begins to form a tiny nucleolus, but these rapidly coalesce as they grow to form the single large nucleolus typical of many interphase cells. (Courtesy of E.G. Jordan and J. McGovern.)

"speckles"). Like the nucleolus, these other nuclear structures lack membranes and are highly dynamic; their appearance is probably the result of the tight association of protein and RNA components involved in the synthesis, assembly, and storage of macromolecules involved in gene expression. Cajal bodies and GEMS resemble one another and are frequently paired in the nucleus; it is not clear whether they truly represent distinct structures. These are likely to be the locations in which snoRNAs and snRNAs undergo covalent modifications and final assembly with proteins. A group of guide RNAs, termed *small Cajal RNAs (scaRNAs)*, selects the sites of these modifications through base pairing. Cajal bodies/GEMS may also be sites where the snRNPs are recycled and their RNAs are "reset" after the rearrangements that occur during splicing (see p. 352). In contrast, the interchromatin granule clusters have been proposed to be stockpiles of fully mature snRNPs and other RNA processing components that are ready to be used in the production of mRNA (**Figure 6–49**).

Scientists have had difficulties in working out the function of these small subnuclear structures, in part because their appearances differ between organisms and can change dramatically as cells traverse the cell cycle or respond to changes in their environment. Much of the progress now being made depends on genetic tools—examination of the effects of designed mutations in model

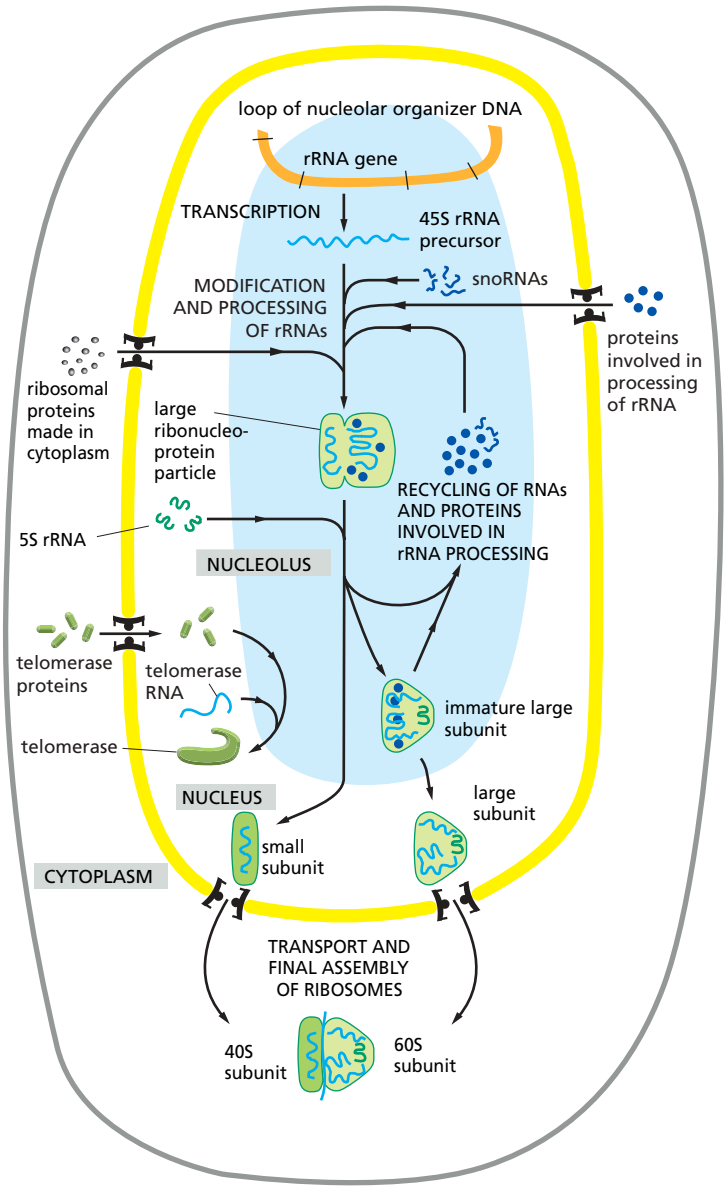


Figure 6–47 The function of the nucleolus in ribosome and other ribonucleoprotein synthesis. The 45S precursor rRNA is packaged in a large ribonucleoprotein particle containing many ribosomal proteins imported from the cytoplasm. While this particle remains at the nucleolus, selected pieces are added and others discarded as it is processed into immature large and small ribosomal subunits. The two ribosomal subunits are thought to attain their final functional form only as each is individually transported through the nuclear pores into the cytoplasm. Other ribonucleoprotein complexes, including telomerase shown here, are also assembled in the nucleolus.

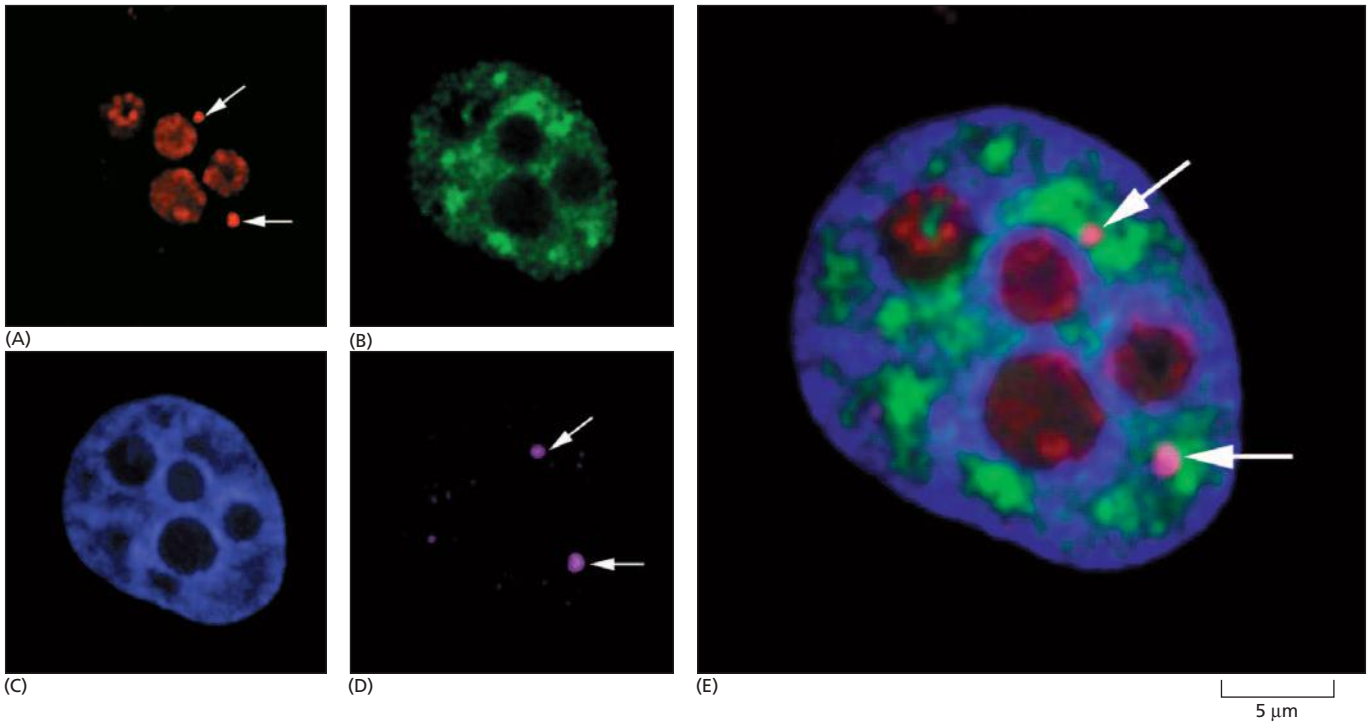


Figure 6–48 Visualization of some prominent nuclear bodies. (A)–(D) Micrographs of the same human cell nucleus, each processed to show a particular set of nuclear structures. (E) All four images enlarged and superimposed. (A) shows the location of the protein fibrillarin (a component of several snoRNPs), which is present at both nucleoli and Cajal bodies, the latter indicated by arrows. (B) shows interchromatin granule clusters or "speckles" detected by using antibodies against a protein involved in pre-mRNA splicing. (C) is stained to show bulk chromatin. (D) shows the location of the protein coilin, which is present at Cajal bodies (*arrows*; see also Figure 4–67). (From J.R. Swedlow and A.I. Lamond, *Gen. Biol.* 2:1–7, 2001. With permission from BioMed Central. Micrographs courtesy of Judith Sleeman.)

organisms or of spontaneous mutations in humans. As one example, GEMS contain the SMN (survival of motor neurons) protein. Certain mutations of the gene encoding this protein are the cause of inherited spinal muscular atrophy, a human disease characterized by a wasting away of the muscles. The disease seems to be caused by a defect in snRNP production. A more complete loss of snRNPs would be expected to be lethal.

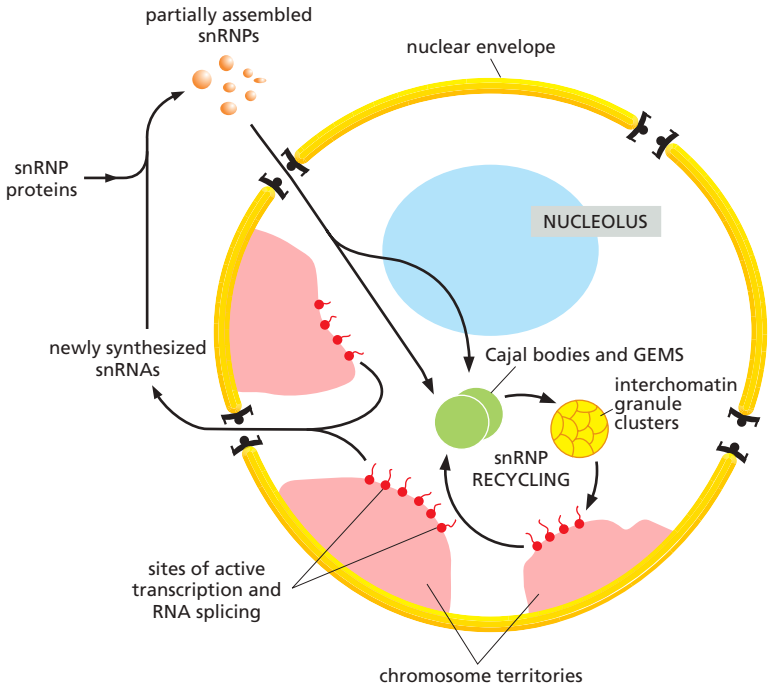


Figure 6–49 Schematic view of subnuclear structures. A typical vertebrate nucleus has several Cajal bodies, which are proposed to be the sites where snRNPs and snoRNPs undergo their final modifications. Interchromatin granule clusters are proposed to be storage sites for fully mature snRNPs. A typical vertebrate nucleus has 20–50 interchromatin granule clusters.

After their initial synthesis, snRNAs are exported from the nucleus to undergo 5′ and 3′ end-processing and assemble with the seven common snRNP proteins (called Sm proteins). These complexes are reimported into the nucleus and the snRNPs undergo their final modification by scaRNAs at Cajal bodies. In addition, snoRNAs chemically modify the U6 snRNP at the nucleolus. The sites of active transcription and splicing (approximately 2000–3000 sites per vertebrate nucleus) correspond to the "perichromatin fibers" seen under the electron microscope. (Adapted from J.D. Lewis and D. Tollervey, *Science* 288:1385–1389, 2000. With permission from AAAS.)

Given the importance of nuclear subdomains in RNA processing, it might have been expected that pre-mRNA splicing would occur in a particular location in the nucleus, as it requires numerous RNA and protein components. However, the assembly of splicing components on pre-mRNA is co-transcriptional; thus, splicing must occur at many locations along chromosomes. Although a typical mammalian cell may be expressing on the order of 15,000 genes, transcription and RNA splicing may be localized to only several thousand sites in the nucleus. These sites themselves are highly dynamic and probably result from the association of transcription and splicing components to create small "assembly lines" with a high local concentration of these components. Interchromatin granule clusters—which contain stockpiles of RNA-processing components—are often observed next to sites of transcription, as though poised to replenish supplies. Thus, the nucleus seems to be highly organized into subdomains, with snRNPs, snoRNPs, and other nuclear components moving between them in an orderly fashion according to the needs of the cell (see Figure 6–48; also see Figure 4–69).

Summary

Before the synthesis of a particular protein can begin, the corresponding mRNA molecule must be produced by transcription. Bacteria contain a single type of RNA polymerase (the enzyme that carries out the transcription of DNA into RNA). An mRNA molecule is produced when this enzyme initiates transcription at a promoter, synthesizes the RNA by chain elongation, stops transcription at a terminator, and releases both the DNA template and the completed mRNA molecule. In eucaryotic cells, the process of transcription is much more complex, and there are three RNA polymerases—polymerase I, II, and III—that are related evolutionarily to one another and to the bacterial polymerase.

RNA polymerase II synthesizes eucaryotic mRNA. This enzyme requires a series of additional proteins, the general transcription factors, to initiate transcription on a purified DNA template, and still more proteins (including chromatin-remodeling complexes and histone-modifying enzymes) to initiate transcription on its chromatin templates inside the cell.

During the elongation phase of transcription, the nascent RNA undergoes three types of processing events: a special nucleotide is added to its 5′ end (capping), intron sequences are removed from the middle of the RNA molecule (splicing), and the 3′ end of the RNA is generated (cleavage and polyadenylation). Each of these processes is initiated by proteins that travel along with RNA polymerase II by binding to sites on its long, extended C-terminal tail. Splicing is unusual in that many of its key steps are carried out by specialized RNA molecules rather than proteins. Properly processed mRNAs are passed through nuclear pore complexes into the cytosol, where they are translated into protein.

For some genes, RNA is the final product. In eucaryotes, these genes are usually transcribed by either RNA polymerase I or RNA polymerase III. RNA polymerase I makes the ribosomal RNAs. After their synthesis as a large precursor, the rRNAs are chemically modified, cleaved, and assembled into the two ribosomal subunits in the nucleolus—a distinct subnuclear structure that also helps to process some smaller RNA–protein complexes in the cell. Additional subnuclear structures (including Cajal bodies and interchromatin granule clusters) are sites where components involved in RNA processing are assembled, stored, and recycled.

FROM RNA TO PROTEIN

In the preceding section we have seen that the final product of some genes is an RNA molecule itself, such as those present in the snRNPs and in ribosomes. However, most genes in a cell produce mRNA molecules that serve as intermediaries on the pathway to proteins. In this section we examine how the cell converts the information carried in an mRNA molecule into a protein molecule. This feat of translation was a focus of attention of biologists in the late 1950s, when it

was posed as the "coding problem": how is the information in a linear sequence of nucleotides in RNA translated into the linear sequence of a chemically quite different set of units—the amino acids in proteins? This fascinating question stimulated great excitement among scientists at the time. Here was a cryptogram set up by nature that, after more than 3 billion years of evolution, could finally be solved by one of the products of evolution—human beings. And indeed, not only has the code been cracked step by step, but in the year 2000 the structure of the elaborate machinery by which cells read this code—the ribosome—was finally revealed in atomic detail.

An mRNA Sequence Is Decoded in Sets of Three Nucleotides

Once an mRNA has been produced by transcription and processing, the information present in its nucleotide sequence is used to synthesize a protein. Transcription is simple to understand as a means of information transfer: since DNA and RNA are chemically and structurally similar, the DNA can act as a direct template for the synthesis of RNA by complementary base-pairing. As the term *transcription* signifies, it is as if a message written out by hand is being converted, say, into a typewritten text. The language itself and the form of the message do not change, and the symbols used are closely related.

In contrast, the conversion of the information in RNA into protein represents a **translation** of the information into another language that uses quite different symbols. Moreover, since there are only 4 different nucleotides in mRNA and 20 different types of amino acids in a protein, this translation cannot be accounted for by a direct one-to-one correspondence between a nucleotide in RNA and an amino acid in protein. The nucleotide sequence of a gene, through the intermediary of mRNA, is translated into the amino acid sequence of a protein by rules that are known as the **genetic code**. This code was deciphered in the early 1960s.

The sequence of nucleotides in the mRNA molecule is read in consecutive groups of three. RNA is a linear polymer of four different nucleotides, so there are $4 \times 4 \times 4 = 64$ possible combinations of three nucleotides: the triplets AAA, AUA, AUG, and so on. However, only 20 different amino acids are commonly found in proteins. Either some nucleotide triplets are never used, or the code is redundant and some amino acids are specified by more than one triplet. The second possibility is, in fact, the correct one, as shown by the completely deciphered genetic code in **Figure 6–50**. Each group of three consecutive nucleotides in RNA is called a **codon**, and each codon specifies either one amino acid or a stop to the translation process.

This genetic code is used universally in all present-day organisms. Although a few slight differences in the code have been found, these are chiefly in the DNA of mitochondria. Mitochondria have their own transcription and protein synthesis systems that operate quite independently from those of the rest of the cell, and it is understandable that their small genomes have been able to accommodate minor changes to the code (discussed in Chapter 14).

	AGA AGG									UUA UUG					AGC AGU				
GCA	CGA						GGA			CUA				CCA	UCA	ACA			GUA	
GCC	CGC						GGC		AUA	CUC				CCC	UCC	ACC			GUC	UAA
GCG	CGG	GAC	AAC	UGC	GAA	CAA	GGG	CAC	AUC	CUG	AAA		UUC	CCG	UCG	ACG		UAC	GUG	UAG
GCU	CGU	GAU	AAU	UGU	GAG	CAG	GGU	CAU	AUU	CUU	AAG	AUG	UUU	CCU	UCU	ACU	UGG	UAU	GUU	UGA
Ala	Arg	Asp	Asn	Cys	Glu	Gln	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val	stop
A	R	D	N	C	E	Q	G	H	I	L	K	M	F	P	S	T	W	Y	V	

Figure 6–50 The genetic code. The standard one-letter abbreviation for each amino acid is presented below its three-letter abbreviation (see Panel 3–1, pp. 128–129, for the full name of each amino acid and its structure). By convention, codons are always written with the 5′-terminal nucleotide to the left. Note that most amino acids are represented by more than one codon, and that there are some regularities in the set of codons that specifies each amino acid. Codons for the same amino acid tend to contain the same nucleotides at the first and second positions, and vary at the third position. Three codons do not specify any amino acid but act as termination sites (stop codons), signaling the end of the protein-coding sequence. One codon—AUG—acts both as an initiation codon, signaling the start of a protein-coding message, and also as the codon that specifies methionine.

In principle, an RNA sequence can be translated in any one of three different **reading frames**, depending on where the decoding process begins (**Figure 6–51**). However, only one of the three possible reading frames in an mRNA encodes the required protein. We see later how a special punctuation signal at the beginning of each RNA message sets the correct reading frame at the start of protein synthesis.

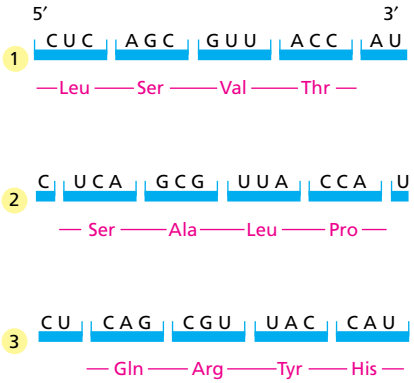


Figure 6–51 The three possible reading frames in protein synthesis. In the process of translating a nucleotide sequence *(blue)* into an amino acid sequence *(red),* the sequence of nucleotides in an mRNA molecule is read from the 5′ end to the 3′ end in consecutive sets of three nucleotides. In principle, therefore, the same RNA sequence can specify three completely different amino acid sequences, depending on the reading frame. In reality, however, only one of these reading frames contains the actual message.

tRNA Molecules Match Amino Acids to Codons in mRNA

The codons in an mRNA molecule do not directly recognize the amino acids they specify: the group of three nucleotides does not, for example, bind directly to the amino acid. Rather, the translation of mRNA into protein depends on adaptor molecules that can recognize and bind both to the codon and, at another site on their surface, to the amino acid. These adaptors consist of a set of small RNA molecules known as **transfer RNAs** (**tRNAs**), each about 80 nucleotides in length.

We saw earlier in this chapter that RNA molecules can fold into precise three-dimensional structures, and the tRNA molecules provide a striking example. Four short segments of the folded tRNA are double-helical, producing a molecule that looks like a cloverleaf when drawn schematically (**Figure 6–52**). For example, a 5′-GCUC-3′ sequence in one part of a polynucleotide chain can form a relatively strong association with a 5′-GAGC-3′ sequence in another region of the same molecule. The cloverleaf undergoes further folding to form a compact L-shaped structure that is held together by additional hydrogen bonds between different regions of the molecule.

Two regions of unpaired nucleotides situated at either end of the L-shaped molecule are crucial to the function of tRNA in protein synthesis. One of these

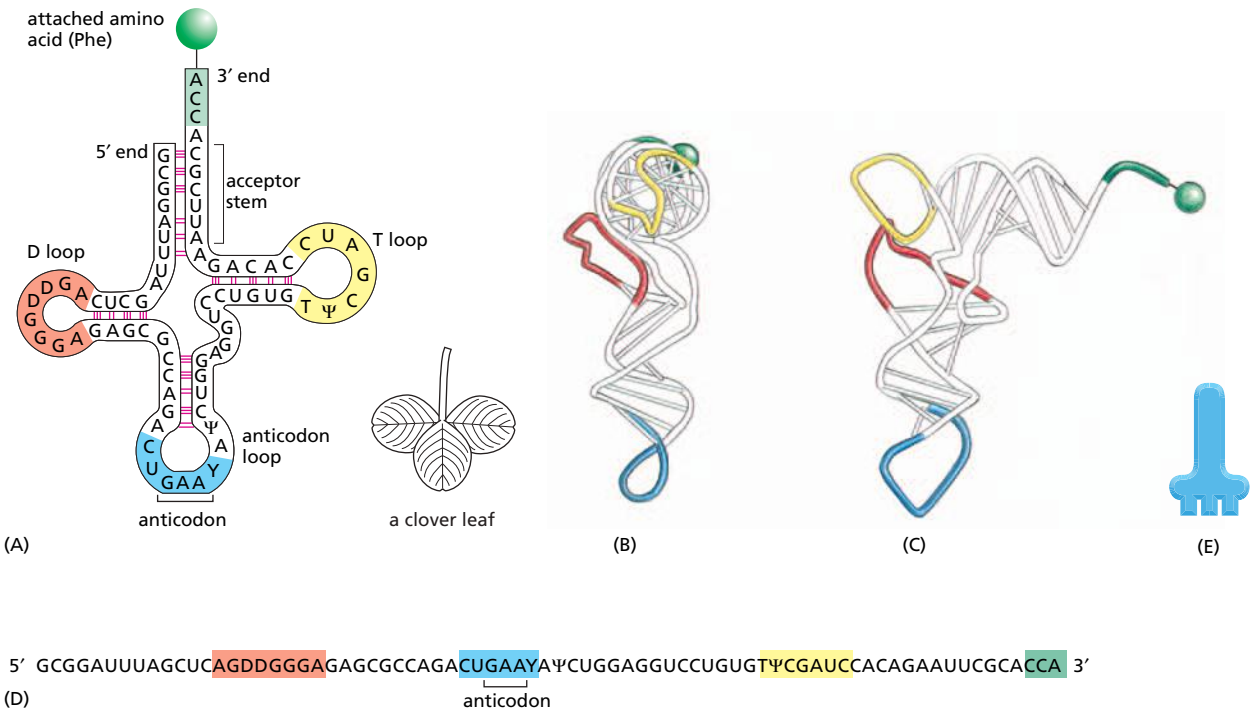


Figure 6–52 A tRNA molecule. A tRNA specific for the amino acid phenylalanine (Phe) is depicted in various ways. (A) The cloverleaf structure showing the complementary base-pairing *(red lines)* that creates the double-helical regions of the molecule. The anticodon is the sequence of three nucleotides that base-pairs with a codon in mRNA. The amino acid matching the codon/anticodon pair is attached at the 3′ end of the tRNA. tRNAs contain some unusual bases, which are produced by chemical modification after the tRNA has been synthesized. For example, the bases denoted ψ (pseudouridine—see Figure 6–43) and D (dihydrouridine—see Figure 6–55) are derived from uracil. (B and C) Views of the L-shaped molecule, based on x-ray diffraction analysis. Although this diagram shows the tRNA for the amino acid phenylalanine, all other tRNAs have similar structures. <CGCA> (D) The linear nucleotide sequence of the molecule, color-coded to match (A), (B), and (C). (E) The tRNA icon we sue in this book.

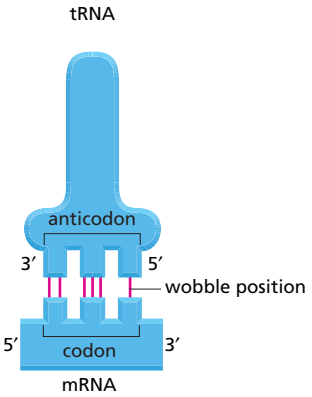
regions forms the **anticodon**, a set of three consecutive nucleotides that pairs with the complementary codon in an mRNA molecule. The other is a short single-stranded region at the 3′ end of the molecule; this is the site where the amino acid that matches the codon is attached to the tRNA.

We have seen in the previous section that the genetic code is redundant; that is, several different codons can specify a single amino acid (see Figure 6–50). This redundancy implies either that there is more than one tRNA for many of the amino acids or that some tRNA molecules can base-pair with more than one codon. In fact, both situations occur. Some amino acids have more than one tRNA and some tRNAs are constructed so that they require accurate base-pairing only at the first two positions of the codon and can tolerate a mismatch (or *wobble*) at the third position (**Figure 6–53**). This wobble base-pairing explains why so many of the alternative codons for an amino acid differ only in their third nucleotide (see Figure 6–50). In bacteria, wobble base-pairings make it possible to fit the 20 amino acids to their 61 codons with as few as 31 kinds of tRNA molecules. The exact number of different kinds of tRNAs, however, differs from one species to the next. For example, humans have nearly 500 tRNA genes but, among them, only 48 different anticodons are represented.

tRNAs Are Covalently Modified Before They Exit from the Nucleus

Like most other eucaryotic RNAs, tRNAs are covalently modified before they are allowed to exit from the nucleus. Eucaryotic tRNAs are synthesized by RNA polymerase III. Both bacterial and eucaryotic tRNAs are typically synthesized as larger precursor tRNAs, which are then trimmed to produce the mature tRNA. In addition, some tRNA precursors (from both bacteria and eucaryotes) contain introns that must be spliced out. This splicing reaction differs chemically from pre-mRNA splicing; rather than generating a lariat intermediate, tRNA splicing uses a cut-and-paste mechanism that is catalyzed by proteins (**Figure 6–54**). Trimming and splicing both require the precursor tRNA to be correctly folded in its cloverleaf configuration. Because misfolded tRNA precursors will not be processed properly, the trimming and splicing reactions are thought to act as quality-control steps in the generation of tRNAs.

All tRNAs are modified chemically—nearly 1 in 10 nucleotides in each mature tRNA molecule is an altered version of a standard G, U, C, or A ribonucleotide. Over 50 different types of tRNA modifications are known; a few are shown in **Figure 6–55**. Some of the modified nucleotides—most notably inosine, produced by the deamination of adenosine—affect the conformation and base-pairing of the anticodon and thereby facilitate the recognition of the appropriate mRNA codon by the tRNA molecule (see Figure 6–53). Others affect the accuracy with which the tRNA is attached to the correct amino acid.



bacteria

wobble codon base	possible anticodon bases
U	A, G, or I
C	G or I
A	U or I
G	C or U

eucaryotes

wobble codon base	possible anticodon bases
U	A, G, or I
C	G or I
A	U
G	C

Figure 6–53 Wobble base-pairing between codons and anticodons. If the nucleotide listed in the first column is present at the third, or wobble, position of the codon, it can base-pair with any of the nucleotides listed in the second column. Thus, for example, when inosine (I) is present in the wobble position of the tRNA anticodon, the tRNA can recognize any one of three different codons in bacteria and either of two codons in eucaryotes. The inosine in tRNAs is formed from the deamination of guanine (see Figure 6–55), a chemical modification that takes place after the tRNA has been synthesized. The nonstandard base pairs, including those made with inosine, are generally weaker than conventional base pairs. Note that codon–anticodon base pairing is more stringent at positions 1 and 2 of the codon: here only conventional base pairs are permitted. The differences in wobble base-pairing interactions between bacteria and eucaryotes presumably result from subtle structural differences between bacterial and eucaryotic ribosomes, the molecular machines that perform protein synthesis. (Adapted from C. Guthrie and J. Abelson, in The Molecular Biology of the Yeast *Saccharomyces*: Metabolism and Gene Expression, pp. 487–528. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press, 1982.)

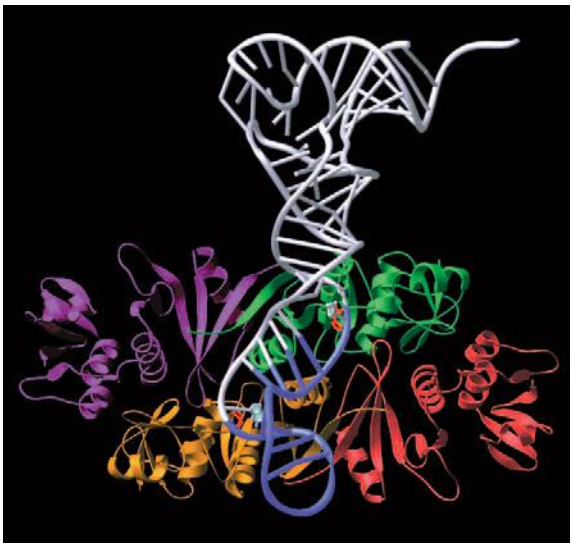


Figure 6–54 Structure of a tRNA-splicing endonuclease docked to a precursor tRNA. The endonuclease (a four-subunit enzyme) removes the tRNA intron *(blue)*. A second enzyme, a multifunctional tRNA ligase (not shown), then joins the two tRNA halves together. (Courtesy of Hong Li, Christopher Trotta and John Abelson.)

Specific Enzymes Couple Each Amino Acid to Its Appropriate tRNA Molecule

We have seen that, to read the genetic code in DNA, cells make a series of different tRNAs. We now consider how each tRNA molecule becomes linked to the one amino acid in 20 that is its appropriate partner. Recognition and attachment of the correct amino acid depends on enzymes called **aminoacyl-tRNA synthetases**, which covalently couple each amino acid to its appropriate set of tRNA molecules (**Figure 6–56** and **Figure 6–57**). Most cells have a different synthetase enzyme for each amino acid (that is, 20 synthetases in all); one attaches glycine to all tRNAs that recognize codons for glycine, another attaches alanine to all tRNAs that recognize codons for alanine, and so on. Many bacteria, however, have fewer than 20 synthetases, and the same synthetase enzyme is responsible for coupling more than one amino acid to the appropriate tRNAs. In these cases, a single synthetase places the identical amino acid on two different types of tRNAs, only one of which has an anticodon that matches the amino acid. A second enzyme then chemically modifies each "incorrectly" attached amino acid so that it now corresponds to the anticodon displayed by its covalently linked tRNA.

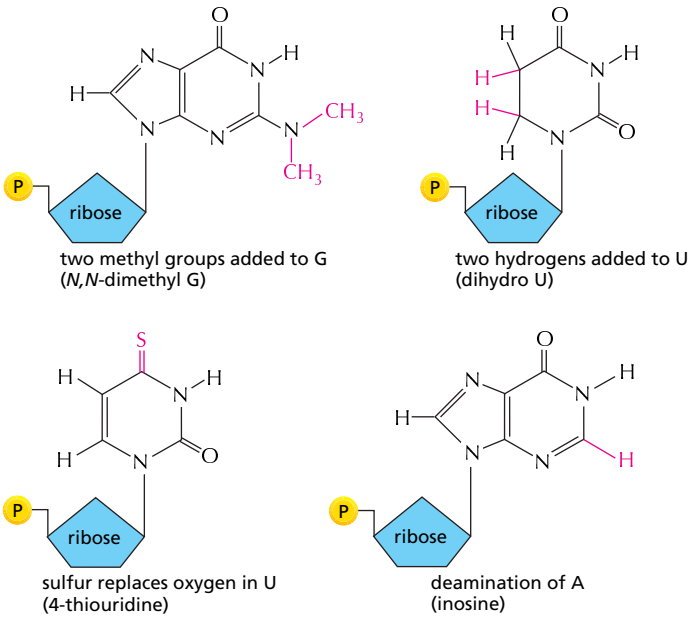


Figure 6–55 A few of the unusual nucleotides found in tRNA molecules. These nucleotides are produced by covalent modification of a normal nucleotide after it has been incorporated into an RNA chain. Two other types of modified nucleotides are shown in Figure 6–43. In most tRNA molecules about 10% of the nucleotides are modified (see Figure 6–52).

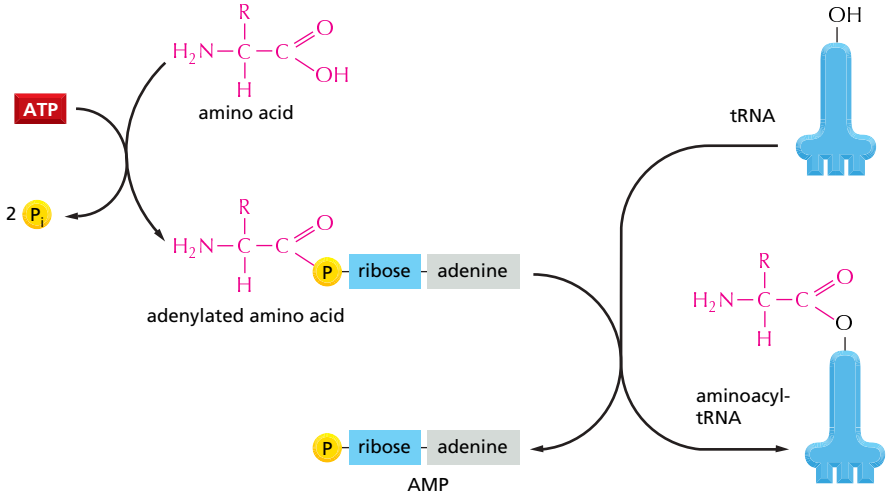


Figure 6–56 Amino acid activation. An amino acid is activated for protein synthesis by an aminoacyl-tRNA synthetase enzyme in two steps. As indicated, the energy of ATP hydrolysis is used to attach each amino acid to its tRNA molecule in a high-energy linkage. The amino acid is first activated through the linkage of its carboxyl group directly to an AMP moiety, forming an *adenylated amino acid*; the linkage of the AMP, normally an unfavorable reaction, is driven by the hydrolysis of the ATP molecule that donates the AMP. Without leaving the synthetase enzyme, the AMP-linked carboxyl group on the amino acid is then transferred to a hydroxyl group on the sugar at the 3′ end of the tRNA molecule. This transfer joins the amino acid by an activated ester linkage to the tRNA and forms the final aminoacyl-tRNA molecule. The synthetase enzyme is not shown in this diagram.

The synthetase-catalyzed reaction that attaches the amino acid to the 3′ end of the tRNA is one of many reactions coupled to the energy-releasing hydrolysis of ATP (see pp. 79–81), and it produces a high-energy bond between the tRNA and the amino acid. The energy of this bond is used at a later stage in protein synthesis to link the amino acid covalently to the growing polypeptide chain.

The aminoacyl-tRNA synthetase enzymes and the tRNAs are equally important in the decoding process (**Figure 6–58**). This was established by an experiment in which one amino acid (cysteine) was chemically converted into a different amino acid (alanine) after it already had been attached to its specific tRNA. When such "hybrid" aminoacyl-tRNA molecules were used for protein synthesis in a cell-free system, the wrong amino acid was inserted at every point in the protein chain where that tRNA was used. Although, as we shall see, cells have several quality control mechanisms to avoid this type of mishap, the experiment establishes that the genetic code is translated by two sets of adaptors that act sequentially. Each matches one molecular surface to another with great specificity, and it is their combined action that associates each sequence of three nucleotides in the mRNA molecule—that is, each codon—with its particular amino acid.

Editing by tRNA Synthetases Ensures Accuracy

Several mechanisms working together ensure that the tRNA synthetase links the correct amino acid to each tRNA. The synthetase must first select the correct amino acid, and most synthetases do so by a two-step mechanism. First, the correct amino acid has the highest affinity for the active-site pocket of its synthetase

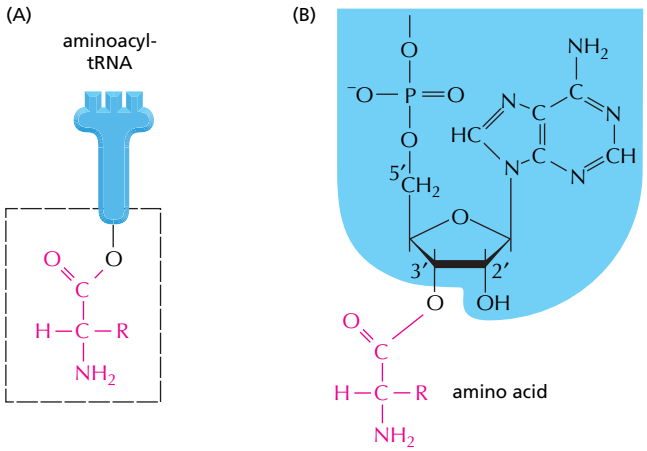
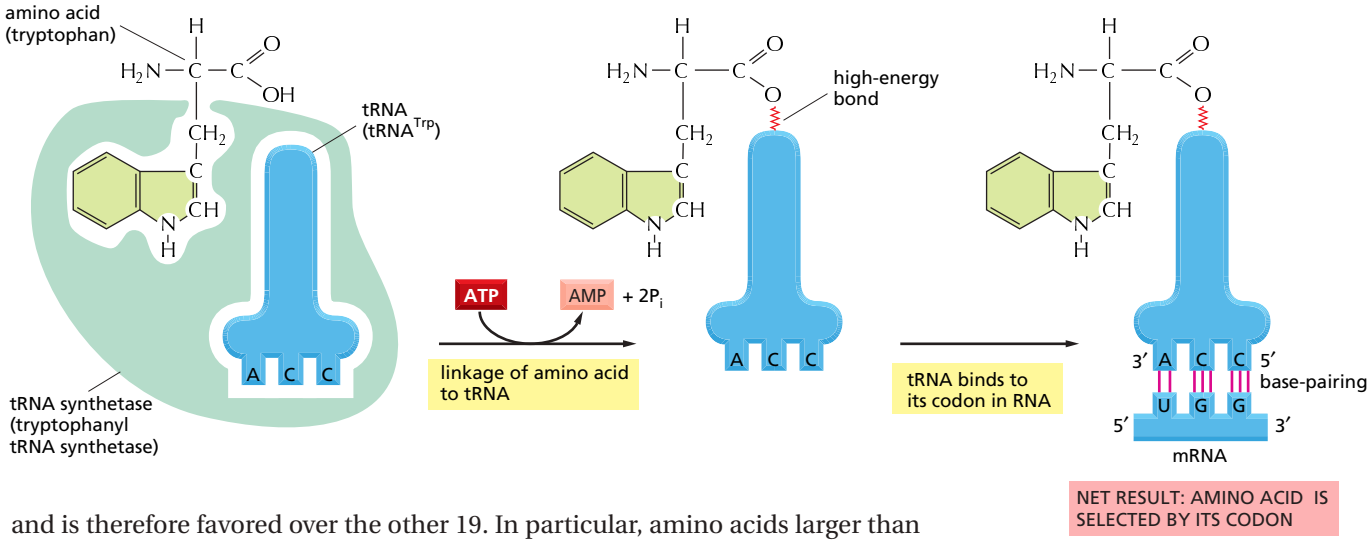


Figure 6–57 The structure of the aminoacyl-tRNA linkage. The carboxyl end of the amino acid forms an ester bond to ribose. Because the hydrolysis of this ester bond is associated with a large favorable change in free energy, an amino acid held in this way is said to be activated. (A) Schematic drawing of the structure. The amino acid is linked to the nucleotide at the 3′ end of the tRNA (see Figure 6–52). (B) Actual structure corresponding to the boxed region in (A). There are two major classes of synthetase enzymes: one links the amino acid directly to the 3′-OH group of the ribose, and the other links it initially to the 2′-OH group. In the latter case, a subsequent transesterification reaction shifts the amino acid to the 3′ position. As in Figure 6–56, the "R group" indicates the side chain of the amino acid.



and is therefore favored over the other 19. In particular, amino acids larger than the correct one are effectively excluded from the active site. However, accurate discrimination between two similar amino acids, such as isoleucine and valine (which differ by only a methyl group), is very difficult to achieve by a one-step recognition mechanism. A second discrimination step occurs after the amino acid has been covalently linked to AMP (see Figure 6–56). When tRNA binds the synthetase, it tries to force the amino acid into a second pocket in the synthetase, the precise dimensions of which exclude the correct amino acid but allow access by closely related amino acids. Once an amino acid enters this editing pocket, it is hydrolyzed from the AMP (or from the tRNA itself if the aminoacyl-tRNA bond has already formed), and is released from the enzyme. This hydrolytic editing, which is analogous to the exonucleolytic proofreading by DNA polymerases (**Figure 6–59**), raises the overall accuracy of tRNA charging to approximately one mistake in 40,000 couplings.

Figure 6–58 The genetic code is translated by means of two adaptors that act one after another. The first adaptor is the aminoacyl-tRNA synthetase, which couples a particular amino acid to its corresponding tRNA; the second adaptor is the tRNA molecule itself, whose *anticodon* forms base pairs with the appropriate *codon* on the mRNA. An error in either step would cause the wrong amino acid to be incorporated into a protein chain. In the sequence of events shown, the amino acid tryptophan (Trp) is selected by the codon UGG on the mRNA.

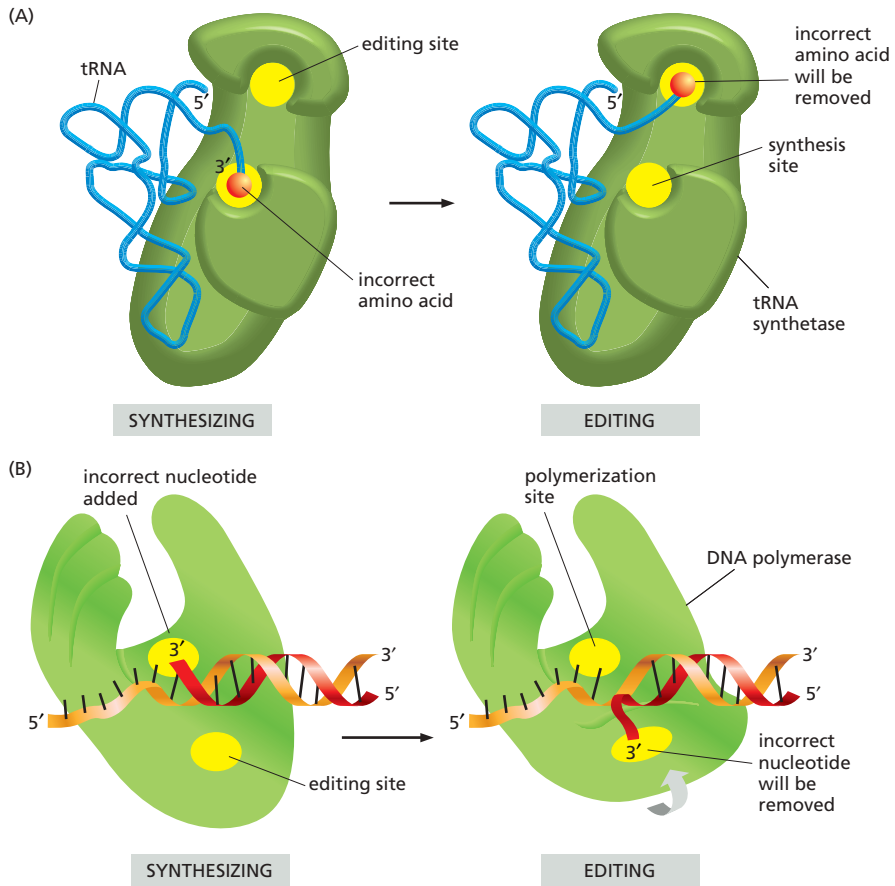


Figure 6–59 Hydrolytic editing. (A) tRNA synthetases remove their own coupling errors through hydrolytic editing of incorrectly attached amino acids. As described in the text, the correct amino acid is rejected by the editing site. (B) The error-correction process performed by DNA polymerase shows some similarities; however, it differs in so far as the removal process depends strongly on a mispairing with the template (see Figure 5–8).

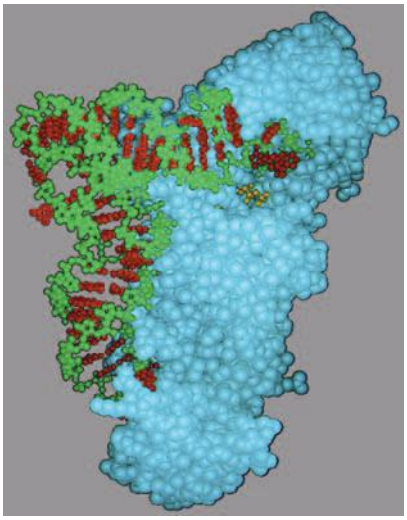


Figure 6–60 The recognition of a tRNA molecule by its aminoacyl-tRNA synthetase. For this tRNA ($tRNA^{Gln}$), specific nucleotides in both the anticodon *(bottom)* and the amino acid-accepting arm allow the correct tRNA to be recognized by the synthetase enzyme *(blue)*. A bound ATP molecule is *yellow*. (Courtesy of Tom Steitz.)

The tRNA synthetase must also recognize the correct set of tRNAs, and extensive structural and chemical complementarity between the synthetase and the tRNA allows the synthetase to probe various features of the tRNA (**Figure 6–60**). Most tRNA synthetases directly recognize the matching tRNA anticodon; these synthetases contain three adjacent nucleotide-binding pockets, each of which is complementary in shape and charge to a nucleotide in the anticodon. For other synthetases, the nucleotide sequence of the acceptor stem is the key recognition determinant. In most cases, however, the synthetase "reads" the nucleotides at several different positions on the tRNA.

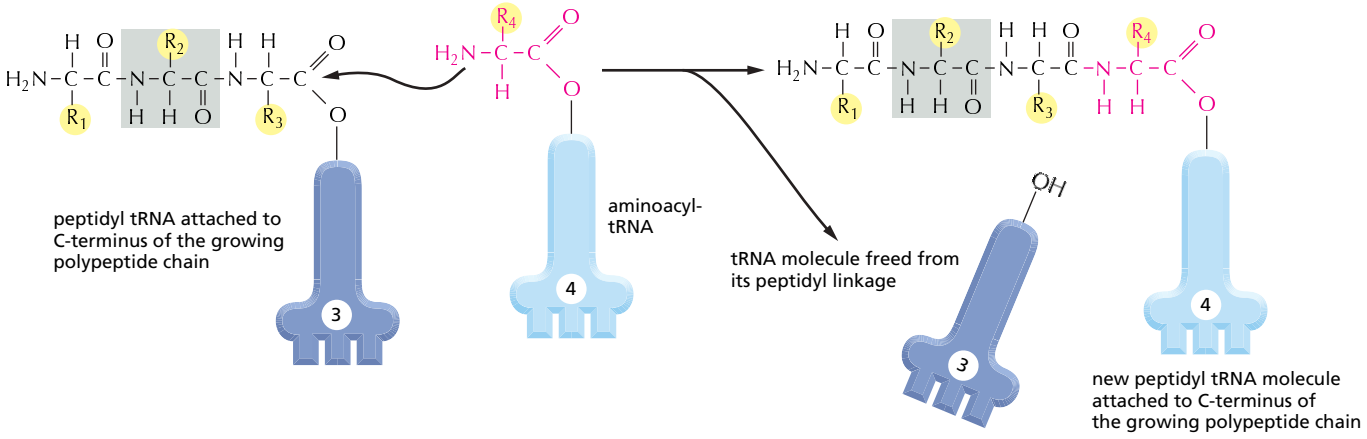
Amino Acids Are Added to the C-terminal End of a Growing Polypeptide Chain

Having seen that amino acids are first coupled to tRNA molecules, we now turn to the mechanism that joins amino acids together to form proteins. The fundamental reaction of protein synthesis is the formation of a peptide bond between the carboxyl group at the end of a growing polypeptide chain and a free amino group on an incoming amino acid. Consequently, a protein is synthesized stepwise from its N-terminal end to its C-terminal end. Throughout the entire process the growing carboxyl end of the polypeptide chain remains activated by its covalent attachment to a tRNA molecule (forming a peptidyl-tRNA). Each addition disrupts this high-energy covalent linkage, but immediately replaces it with an identical linkage on the most recently added amino acid (**Figure 6–61**). In this way, each amino acid added carries with it the activation energy for the addition of the next amino acid rather than the energy for its own addition—an example of the "head growth" type of polymerization described in Figure 2–68.

The RNA Message Is Decoded in Ribosomes

The synthesis of proteins is guided by information carried by mRNA molecules. To maintain the correct reading frame and to ensure accuracy (about 1 mistake every 10,000 amino acids), protein synthesis is performed in the **ribosome**, a complex catalytic machine made from more than 50 different proteins (the *ribosomal proteins*) and several RNA molecules, the **ribosomal RNAs** (**rRNAs**). <CGCC> A typical eucaryotic cell contains millions of ribosomes in its cytoplasm (**Figure 6–62**). Eucaryotic ribosome subunits are assembled at the nucleolus, when newly transcribed and modified rRNAs associate with ribosomal

Figure 6–61 The incorporation of an amino acid into a protein. A polypeptide chain grows by the stepwise addition of amino acids to its C-terminal end. The formation of each peptide bond is energetically favorable because the growing C-terminus has been activated by the covalent attachment of a tRNA molecule. The peptidyl-tRNA linkage that activates the growing end is regenerated during each addition. The amino acid side chains have been abbreviated as R_1, R_2, R_3, and R_4; as a reference point, all of the atoms in the second amino acid in the polypeptide chain are shaded *gray*. The figure shows the addition of the fourth amino acid *(red)* to the growing chain.



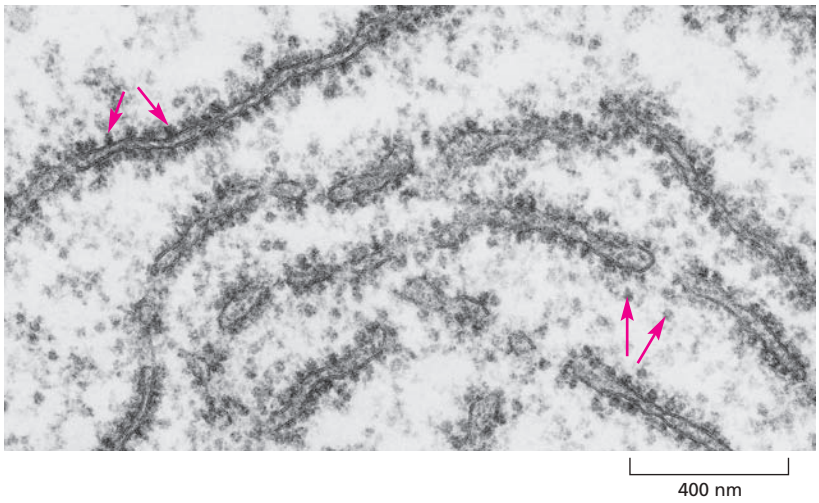


Figure 6–62 Ribosomes in the cytoplasm of a eucaryotic cell. This electron micrograph shows a thin section of a small region of cytoplasm. The ribosomes appear as black dots *(red arrows)*. Some are free in the cytosol; others are attached to membranes of the endoplasmic reticulum. (Courtesy of Daniel S. Friend.)

proteins, which have been transported into the nucleus after their synthesis in the cytoplasm. The two ribosomal subunits are then exported to the cytoplasm, where they join together to synthesize proteins.

Eucaryotic and procaryotic ribosomes have similar designs and functions. Both are composed of one large and one small subunit that fit together to form a complete ribosome with a mass of several million daltons (**Figure 6–63**). The small subunit provides the framework on which the tRNAs can be accurately

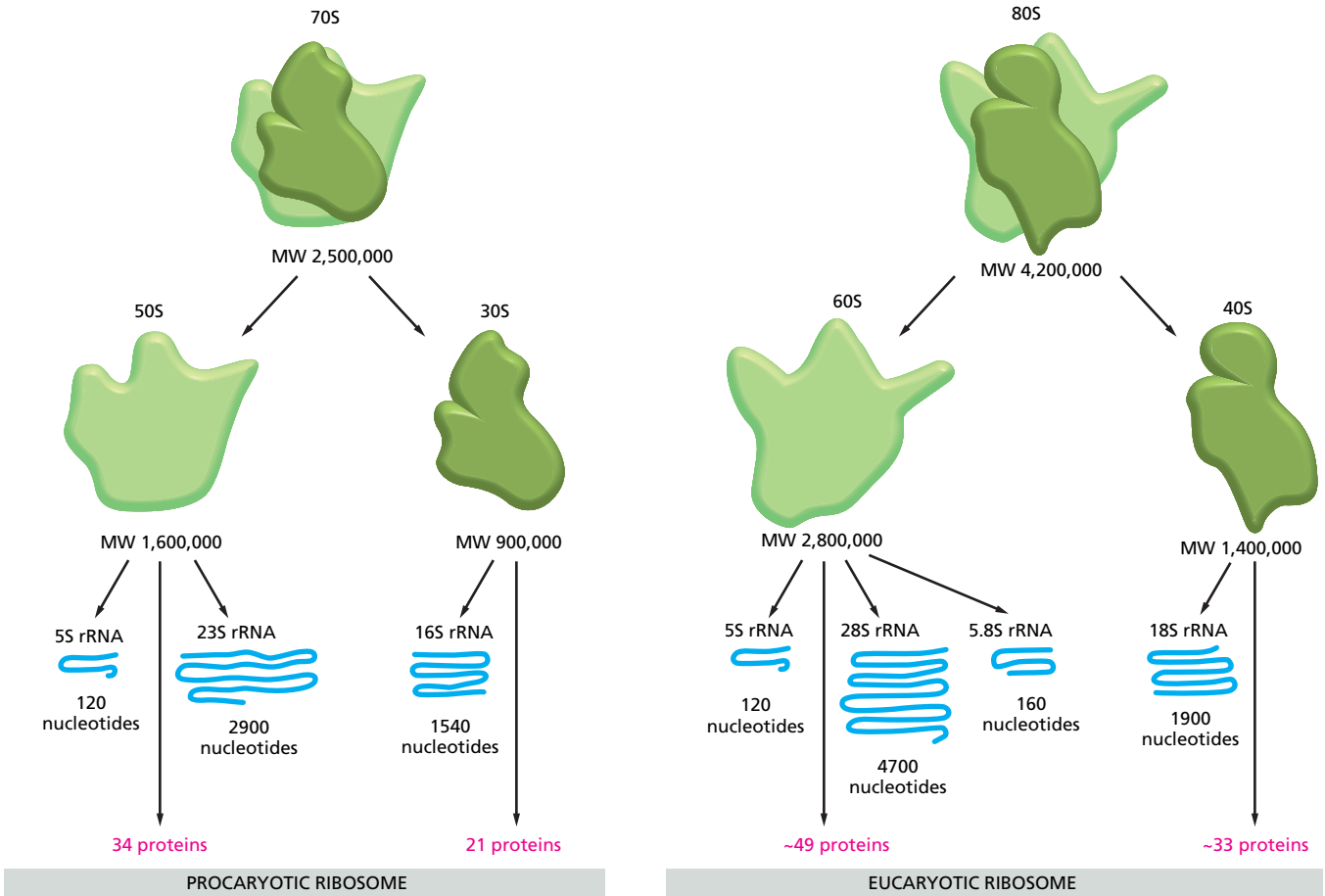


Figure 6–63 A comparison of procaryotic and eucaryotic ribosomes. Despite differences in the number and size of their rRNA and protein components, both procaryotic and eucaryotic ribosomes have nearly the same structure and they function similarly. Although the 18S and 28S rRNAs of the eucaryotic ribosome contain many nucleotides not present in their bacterial counterparts, these nucleotides are present as multiple insertions that form extra domains and leave the basic structure of each rRNA largely unchanged.

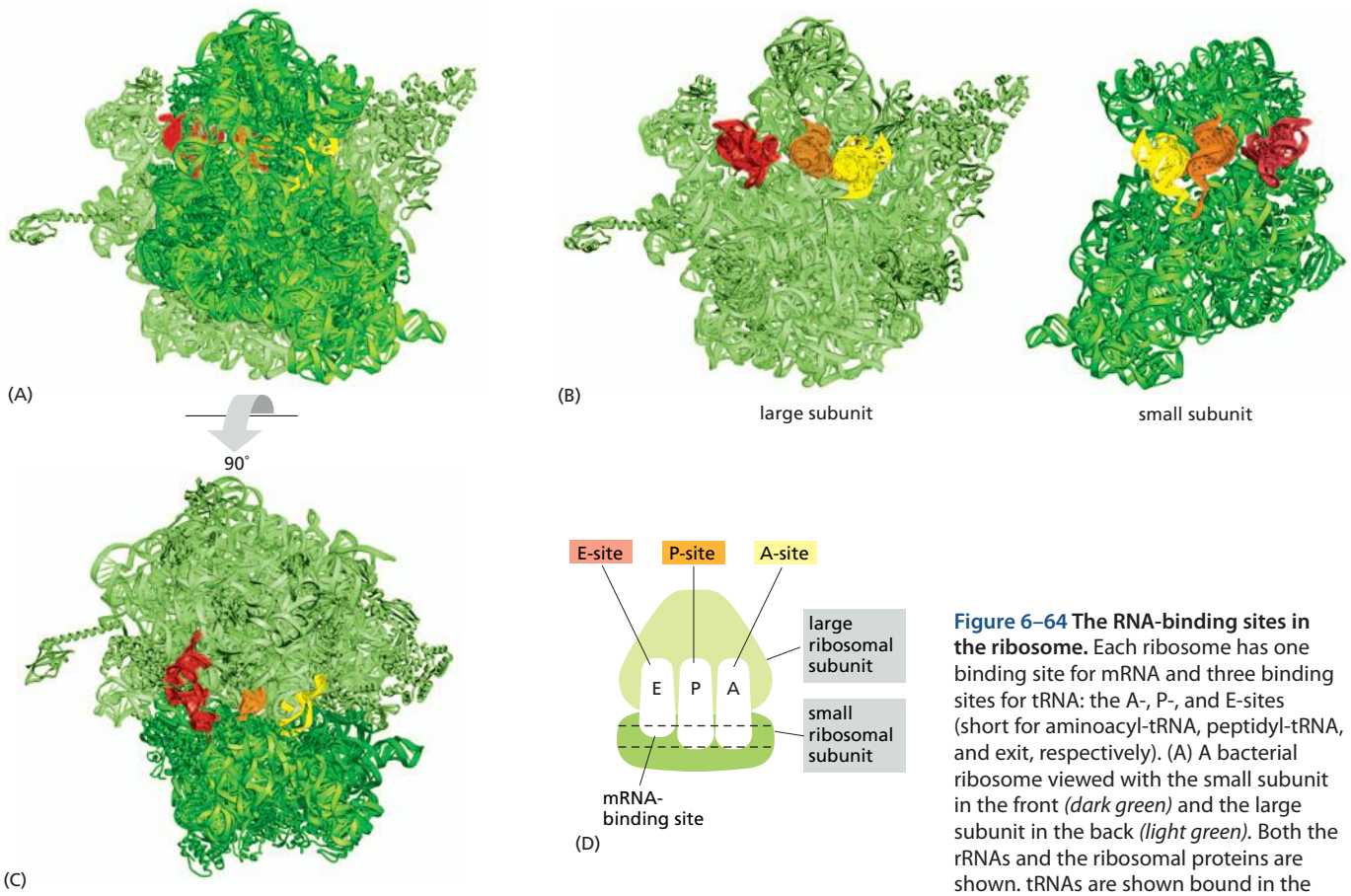


Figure 6–64 The RNA-binding sites in the ribosome. Each ribosome has one binding site for mRNA and three binding sites for tRNA: the A-, P-, and E-sites (short for aminoacyl-tRNA, peptidyl-tRNA, and exit, respectively). (A) A bacterial ribosome viewed with the small subunit in the front *(dark green)* and the large subunit in the back *(light green)*. Both the rRNAs and the ribosomal proteins are shown. tRNAs are shown bound in the E-site *(red)*, the P-site *(orange)* and the A-site *(yellow)*. Although all three tRNA sites are shown occupied here, during the process of protein synthesis not more than two of these sites are thought to contain tRNA molecules at any one time (see Figure 6–66). (B) Large and small ribosomal subunits arranged as though the ribosome in (A) were opened like a book. (C) The ribosome in (A) rotated through 90º and viewed with the large subunit on top and small subunit on the bottom. (D) Schematic representation of a ribosome (in the same orientation as C), which will be used in subsequent figures. (A, B, and C, adapted from M.M. Yusupov et al., *Science* 292:883–896, 2001. With permission from AAAS; courtesy of Albion Baucom and Harry Noller.)

matched to the codons of the mRNA (see Figure 6–58), while the large subunit catalyzes the formation of the peptide bonds that link the amino acids together into a polypeptide chain (see Figure 6–61).

When not actively synthesizing proteins, the two subunits of the ribosome are separate. They join together on an mRNA molecule, usually near its 5′ end, to initiate the synthesis of a protein. The mRNA is then pulled through the ribosome; as its codons enter the core of the ribosome, the mRNA nucleotide sequence is translated into an amino acid sequence using the tRNAs as adaptors to add each amino acid in the correct sequence to the end of the growing polypeptide chain. When a stop codon is encountered, the ribosome releases the finished protein, and its two subunits separate again. These subunits can then be used to start the synthesis of another protein on another mRNA molecule.

Ribosomes operate with remarkable efficiency: in one second, a single ribosome of a eucaryotic cell adds about 2 amino acids to a polypeptide chain; the ribosomes of bacterial cells operate even faster, at a rate of about 20 amino acids per second. How does the ribosome choreograph the many coordinated movements required for efficient translation? A ribosome contains four binding sites for RNA molecules: one is for the mRNA and three (called the A-site, the P-site, and the E-site) are for tRNAs (**Figure 6–64**). A tRNA molecule is held tightly at the A- and P-sites only if its anticodon forms base pairs with a complementary codon (allowing for wobble) on the mRNA molecule that is threaded through the ribosome (**Figure 6–65**). The A- and P-sites are close enough together for their two tRNA molecules to be forced to form base pairs with adjacent codons on the mRNA molecule. This feature of the ribosome maintains the correct reading frame on the mRNA.

Once protein synthesis has been initiated, each new amino acid is added to the elongating chain in a cycle of reactions containing four major steps: tRNA

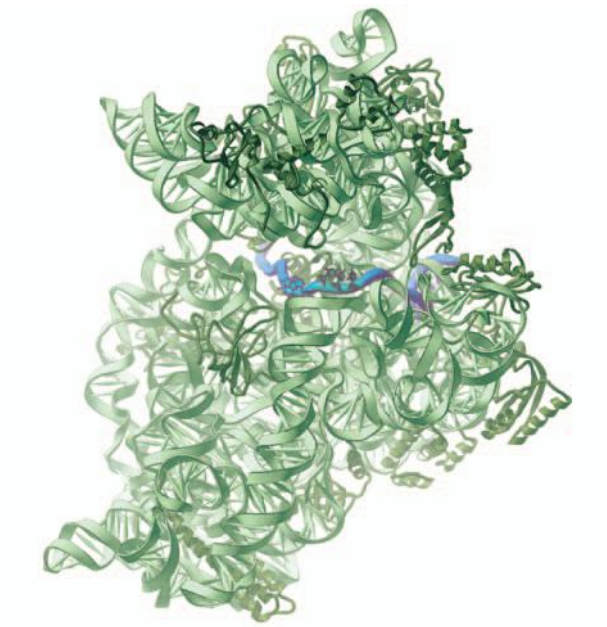


Figure 6–65 The path of mRNA *(blue)* through the small ribosomal subunit. The orientation is the same as that in the right-hand panel of Figure 6–64B. (Courtesy of Harry F. Noller, based on data in G.Z. Yusopova et al., *Cell* 106:233–241, 2001. With permission from Elsevier.)

binding, peptide bond formation, large subunit and small subunit translocation. As a result of the two translocation steps, the entire ribosome moves three nucleotides along the mRNA and is positioned to start the next cycle. (**Figure 6–66**). Our description of the chain elongation process begins at a point at which some amino acids have already been linked together and there is a tRNA molecule in the P-site on the ribosome, covalently joined to the end of the growing polypeptide. In step 1, a tRNA carrying the next amino acid in the chain binds to the ribosomal A-site by forming base pairs with the mRNA codon positioned there, so that the P-site and the A-site contain adjacent bound tRNAs. In step 2, the carboxyl end of the polypeptide chain is released from the tRNA at the P-site (by breakage of the high-energy bond between the tRNA and its amino acid) and joined to the free amino group of the amino acid linked to the tRNA at the A-site, forming a new peptide bond. This central reaction of protein synthesis is catalyzed by a *peptidyl transferase* contained in the large ribosomal subunit. In step 3, the large subunit moves relative to the mRNA held by the small subunit, thereby shifting the acceptor stems of the two tRNAs to the E- and P-sites of the large subunit. In step 4, another series of conformational changes moves the small subunit and its bound mRNA exactly three nucleotides, resetting the ribosome so it is ready to receive the next aminoacyl-tRNA. Step 1 is then repeated with a new incoming aminoacyl-tRNA, and so on. <CGTT>

This four-step cycle is repeated each time an amino acid is added to the polypeptide chain, as the chain grows from its amino to its carboxyl end.

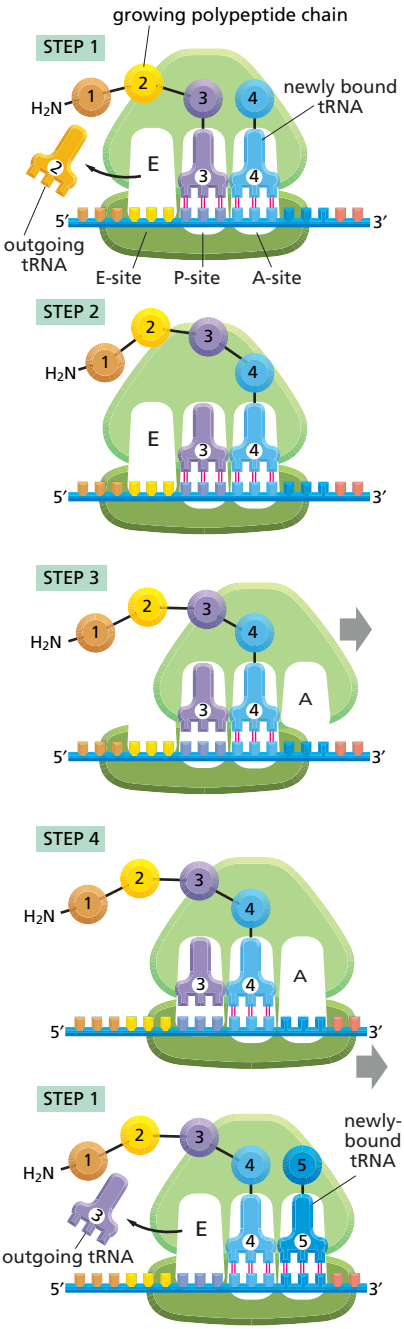


Figure 6–66 Translating an mRNA molecule. Each amino acid added to the growing end of a polypeptide chain is selected by complementary base-pairing between the anticodon on its attached tRNA molecule and the next codon on the mRNA chain. Because only one of the many types of tRNA molecules in a cell can base-pair with each codon, the codon determines the specific amino acid to be added to the growing polypeptide chain. The four-step cycle shown is repeated over and over during the synthesis of a protein. In step 1, an aminoacyl-tRNA molecule binds to a vacant A-site on the ribosome and a spent tRNA molecule dissociates from the E-site. In step 2, a new peptide bond is formed. In step 3, the large subunit translocates relative to the small subunit, leaving the two tRNAs in hybrid sites: P on the large subunit and A on the small, for one; E on the large subunit and P on the small, for the other. In step 4, the small subunit translocates carrying its mRNA a distance of three nucleotides through the ribosome. This "resets" the ribosome with a fully empty A-site, ready for the next aminoacyl-tRNA molecule to bind. As indicated, the mRNA is translated in the 5′-to-3′ direction, and the N-terminal end of a protein is made first, with each cycle adding one amino acid to the C-terminus of the polypeptide chain.

Elongation Factors Drive Translation Forward and Improve Its Accuracy

The basic cycle of polypeptide elongation shown in outline in Figure 6–66 has an additional feature that makes translation especially efficient and accurate. Two *elongation factors* enter and leave the ribosome during each cycle, each hydrolyzing GTP to GDP and undergoing conformational changes in the process. These factors are called EF-Tu and EF-G in bacteria, and EF1 and EF2 in eucaryotes. Under some conditions *in vitro,* ribosomes can be forced to synthesize proteins without the aid of these elongation factors and GTP hydrolysis, but this synthesis is very slow, inefficient, and inaccurate. Coupling the GTP hydrolysis-driven changes in the elongation factors to transitions between different states of the ribosome speeds up protein synthesis enormously. Although these ribosomal states are not yet understood in detail, they almost certainly involve RNA structure rearrangements in the ribosome core. The cycles of elongation factor association, GTP hydrolysis, and dissociation ensure that all such changes occur in the "forward" direction so that translation can proceed efficiently (**Figure 6–67**).

As shown previously, EF-Tu simultaneously binds GTP and aminoacyl-tRNAs (see Figure 3–74). In addition to helping move translation forward, EF-Tu (EF1 in eucaryotes) increases the accuracy of translation in several ways. First, as it escorts an incoming aminoacyl-tRNA to the ribosome, EF-Tu checks whether the tRNA–amino acid match is corrrect. Exactly how this is accomplished is not well understood. According to one idea, correct tRNA–amino acid matches have a narrowly defined affinity for EF-Tu, which allows EF-Tu to discriminate, albeit crudely, among many different amino acid–tRNA combinations, selectively bringing the correct ones with it into the ribosome. Second, EF-Tu monitors the initial interaction between the anticodon of an incoming aminoacyl-tRNA and the codon of the mRNA in the A-site. Aminoacyl-tRNAs are "bent" when bound to the GTP-form of EF-Tu; this bent conformation allows codon pairing but prevents incorporation of the amino acid into the growing polypeptide chain. However, if the codon–anticodon match is correct, the ribosome rapidly triggers the hydrolysis of the GTP molecule, whereupon EF-Tu releases its grip on the tRNA and dissociates from the ribosome, allowing the tRNA to donate its amino acid for protein synthesis. But how is the "correctness" of the codon–anticodon match assessed? This feat is carried out by the ribosome itself through an RNA-based mechanism. The rRNA in the small subunit of the ribosome forms a series of hydrogen bonds with the codon–anticodon pair that allows determination of its correctness (**Figure 6–68**). In essence, the rRNA folds around the codon–anticodon pair, and its final closure—which occurs only when the correct anticodon is in place—triggers GTP hydrolysis. Remarkably, this induced fit mechanism can distinguish correct from incorrect codon–anticodon interactions despite the rules for wobble base-pairing summarized in Figure 6–53. From this example, as for RNA splicing, one gets a sense of the highly sophisticated forms of molecular recognition that can be achieved solely by RNA.

The interactions of EF-Tu, tRNA, and the ribosome just described introduce critical proofreading steps into protein synthesis at the initial tRNA selection stage. But after GTP is hydrolyzed and EF-Tu dissociates from the ribosome, there is an additional opportunity for the ribosome to prevent an incorrect amino acid from being added to the growing chain. Following GTP hydrolysis, there is a short time delay as the amino acid carried by the tRNA moves into position on the ribosome. This time delay is shorter for correct than incorrect codon–anticodon pairs. Moreover, incorrectly matched tRNAs dissociate more

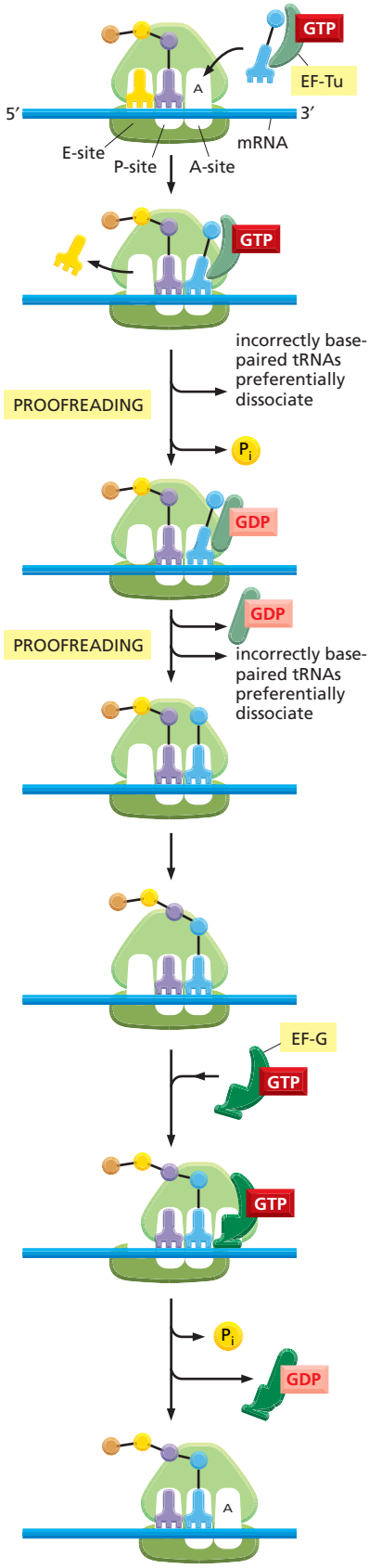


Figure 6–67 Detailed view of the translation cycle. The outline of translation presented in Figure 6–66 has been expanded to show the roles of two elongation factors EF-Tu and EF-G, which drive translation in the forward direction. As explained in the text, EF-Tu also provides two opportunities for proofreading of the codon–anticodon match. In this way, incorrectly paired tRNAs are selectively rejected, and the accuracy of translation is improved.

rapidly than those correctly bound because their interaction with the codon is weaker. Thus, most incorrectly bound tRNA molecules (as well as a significant number of correctly bound molecules) will leave the ribosome without being used for protein synthesis. All of these proofreading steps, taken together, are largely responsible for the 99.99% accuracy of the ribosome in translating RNA into protein.

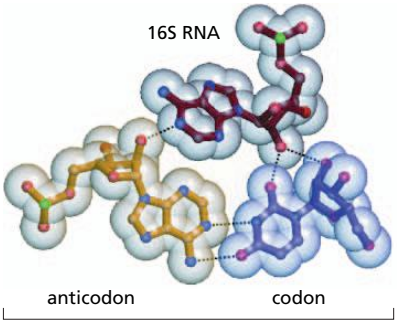


Figure 6–68 Recognition of correct codon–anticodon matches by the small subunit rRNA of the ribosome. Shown is the interaction between a nucleotide of the small subunit rRNA and the first nucleotide pair of a correctly paired codon–anticodon; similar interactions occur between other nucleotides of the rRNA and the second and third positions of the codon–anticodon pair. The small-subunit rRNA can form this network of hydrogen bonds only with correctly matched codon–anticodon pairs. As explained in the text, this codon–anticodon monitoring by the small-subunit rRNA increases the accuracy of protein synthesis. (From J.M. Ogle et al., *Science* 292:897–902, 2001. With permission from AAAS.)

The Ribosome Is a Ribozyme

The ribosome is a large complex composed of two-thirds RNA and one-third protein. The determination, in 2000, of the entire three-dimensional conformation of its large and small subunits is a major triumph of modern structural biology. The findings confirm earlier evidence that rRNAs—and not proteins—are responsible for the ribosome's overall structure, its ability to position tRNAs on the mRNA, and its catalytic activity in forming covalent peptide bonds. The ribosomal RNAs are folded into highly compact, precise three-dimensional structures that form the compact core of the ribosome and determine its overall shape (**Figure 6–69**).

In marked contrast to the central positions of the rRNAs, the ribosomal proteins are generally located on the surface and fill in the gaps and crevices of the folded RNA (**Figure 6–70**). Some of these proteins send out extended regions of polypeptide chain that penetrate short distances into holes in the RNA core (**Figure 6–71**). The main role of the ribosomal proteins seems to be to stabilize the RNA core, while permitting the changes in rRNA conformation that are necessary for this RNA to catalyze efficient protein synthesis. The proteins probably also aid in the initial assembly of the rRNAs that make up the core of the ribosome.

Not only are the A-, P-, and E-binding sites for tRNAs formed primarily by ribosomal RNAs, but the catalytic site for peptide bond formation is also formed by RNA, as the nearest amino acid is located more than 1.8 nm away.

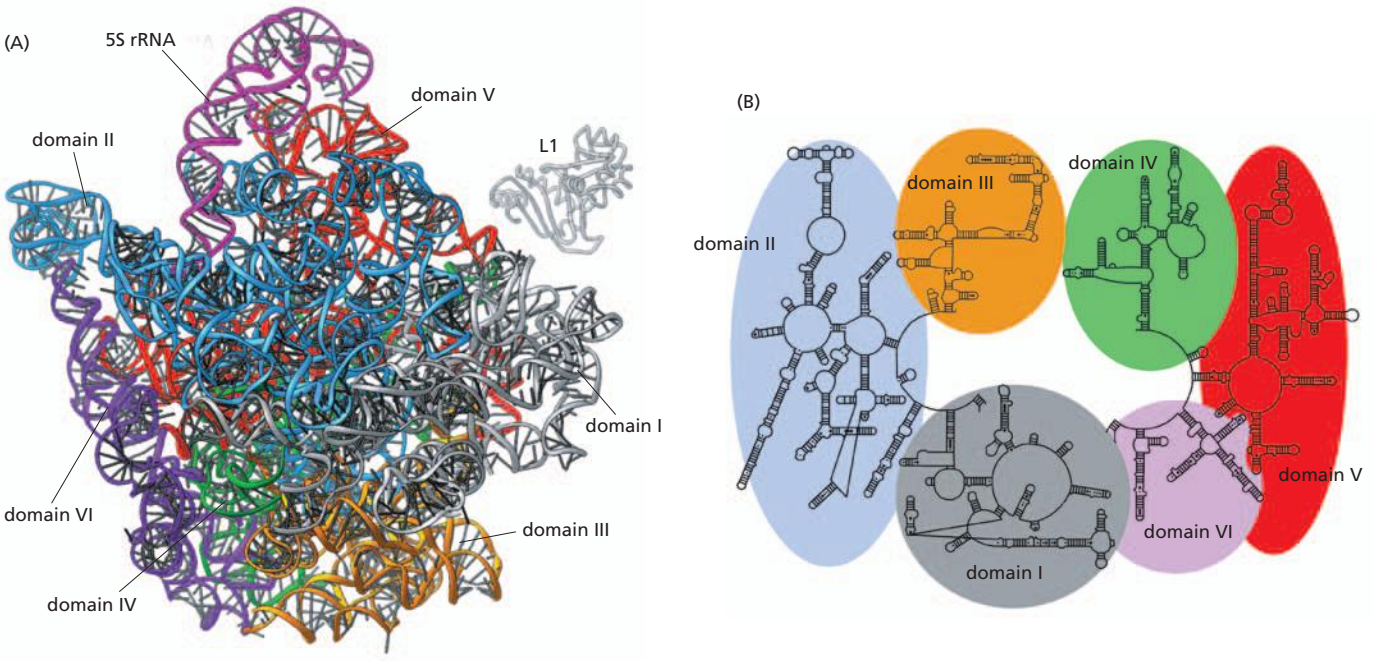


Figure 6–69 Structure of the rRNAs in the large subunit of a bacterial ribosome, as determined by x-ray crystallography. (A) Three-dimensional conformations of the large-subunit rRNAs (5S and 23S) as they appear in the ribosome. One of the protein subunits of the ribosome (L1) is also shown as a reference point, since it forms a characteristic protrusion on the ribosome. (B) Schematic diagram of the secondary structure of the 23S rRNA, showing the extensive network of base-pairing. The structure has been divided into six "domains" whose colors correspond to those in (A). The secondary-structure diagram is highly schematized to represent as much of the structure as possible in two dimensions. To do this, several discontinuities in the RNA chain have been introduced, although in reality the 23S RNA is a single RNA molecule. For example, the base of Domain III is continuous with the base of Domain IV even though a gap appears in the diagram. (Adapted from N. Ban et al., *Science* 289:905–920, 2000. With permission from AAAS.)

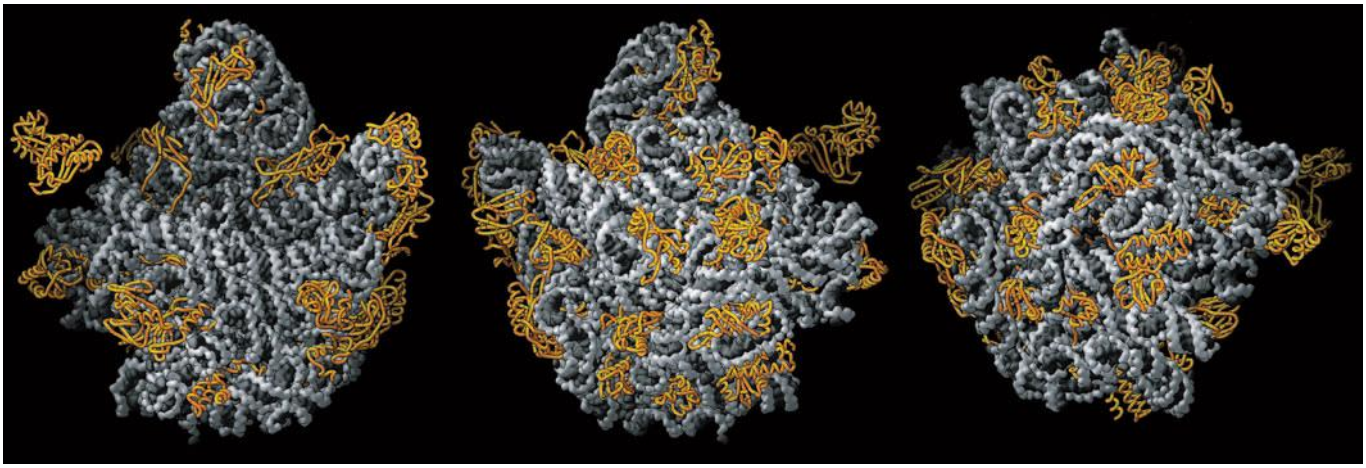


Figure 6–70 Location of the protein components of the bacterial large ribosomal subunit. The rRNAs (5S and 23S) are shown in *gray* and the large-subunit proteins (27 of the 31 total) in *gold*. For convenience, the protein structures depict only the polypeptide backbones. (A) Interface with the small subunit, the same view shown in Figure 6–64B. (B) Side opposite to that shown in (A), obtained by rotating (A) by 180° around a vertical axis. (C) Further slight rotation of (B) through a diagonal axis, allowing a view into the peptide exit channel in the center of the structure. (From N. Ban et al., *Science* 289:905–920, 2000. With permission from AAAS.)

This discovery came as a surprise to biologists because, unlike proteins, RNA does not contain easily ionizable functional groups that can be used to catalyze sophisticated reactions like peptide bond formation. Moreover, metal ions, which are often used by RNA molecules to catalyze chemical reactions (as discussed later in the chapter), were not observed at the active site of the ribosome. Instead, it is believed that the 23S rRNA forms a highly structured pocket that, through a network of hydrogen bonds, precisely orients the two reactants (the growing peptide chain and an aminoacyl-tRNA) and thereby greatly accelerates their covalent joining. In addition, the tRNA in the P site contributes to the active site, perhaps supplying a functional OH group that participates directly in the catalysis. This mechanism may ensure that catalysis occurs only when the tRNA is properly positioned in the ribosome.

RNA molecules that possess catalytic activity are known as **ribozymes**. We saw earlier in this chapter how other ribozymes function in self-splicing reactions (for example, see Figure 6–36). In the final section of this chapter, we consider what the ability of RNA molecules to function as catalysts for a wide variety of different reactions might mean for the early evolution of living cells. For now, we merely note that there is good reason to suspect that RNA rather than protein molecules served as the first catalysts for living cells. If so, the ribosome, with its RNA core, may be a relic of an earlier time in life's history—when protein synthesis evolved in cells that were run almost entirely by ribozymes.

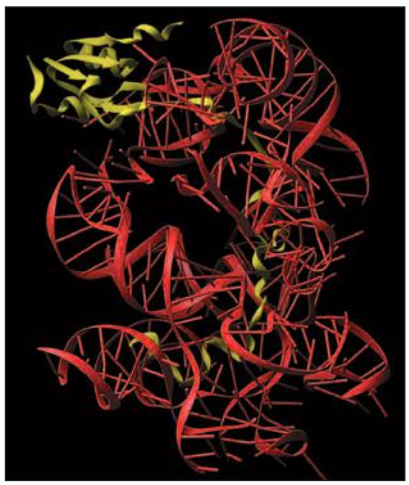


Figure 6–71 Structure of the L15 protein in the large subunit of the bacterial ribosome. The globular domain of the protein lies on the surface of the ribosome and an extended region penetrates deeply into the RNA core of the ribosome. The L15 protein is shown in *yellow* and a portion of the ribosomal RNA core is shown in *red*. (From D. Klein, P.B. Moore and T.A. Steitz, *J. Mol. Biol.* 340:141–147, 2004. With permission from Academic Press.)

Nucleotide Sequences in mRNA Signal Where to Start Protein Synthesis

The initiation and termination of translation share features of the translation elongation cycle described above. The site at which protein synthesis begins on the mRNA is especially crucial, since it sets the reading frame for the whole length of the message. An error of one nucleotide either way at this stage would cause every subsequent codon in the message to be misread, resulting in a nonfunctional protein with a garbled sequence of amino acids. The initiation step is also important because for most genes it is the last point at which the cell can decide whether the mRNA is to be translated and the protein synthesized; the rate of initiation is thus one determinant of the rate at which any protein is synthesized. We shall see in Chapter 7 that cells use several mechanisms to regulate translation initiation.

The translation of an mRNA begins with the codon AUG, and a special tRNA is required to start translation. This **initiator tRNA** always carries the amino acid methionine (in bacteria, a modified form of methionine—formylmethionine—is used), with the result that all newly made proteins have methionine as the first amino acid at their N-terminus, the end of a protein that is synthesized first. This methionine is usually removed later by a specific protease. The initiator tRNA can be specially recognized by initiation factors because it has a nucleotide sequence distinct from that of the tRNA that normally carries methionine.

In eucaryotes, the initiator tRNA–methionine complex (Met–tRNAi) is first loaded into the small ribosomal subunit along with additional proteins called **eucaryotic initiation factors**, or **eIFs** (**Figure 6–72**). Of all the aminoacyl-tRNAs in the cell, only the methionine-charged initiator tRNA is capable of tightly binding the small ribosome subunit without the complete ribosome being present and it binds directly to the P-site. Next, the small ribosomal subunit binds to the 5′ end of an mRNA molecule, which is recognized by virtue of its 5′ cap and its two bound initiation factors, eIF4E (which directly binds the cap) and eIF4G (see Figure 6–40). The small ribosomal subunit then moves forward (5′ to 3′) along the mRNA, searching for the first AUG. Additional initiation factors that act as ATP-powered helicases facilitate the ribosome's movement through RNA secondary structure. In 90% of mRNAs, translation begins at the first AUG encountered by the small subunit. At this point, the initiation factors dissociate, allowing the large ribosomal subunit to assemble with the complex and complete the ribosome. The initiator tRNA is still bound to the P-site, leaving the A-site vacant. Protein synthesis is therefore ready to begin (see Figure 6–72).

The nucleotides immediately surrounding the start site in eucaryotic mRNAs influence the efficiency of AUG recognition during the above scanning process. If this recognition site differs substantially from the consensus recognition sequence (5′-ACCAUGG-3′), scanning ribosomal subunits will sometimes ignore the first AUG codon in the mRNA and skip to the second or third AUG codon instead. Cells frequently use this phenomenon, known as "leaky scanning," to produce two or more proteins, differing in their N-termini, from the same mRNA molecule. It allows some genes to produce the same protein with and without a signal sequence attached at its N-terminus, for example, so that the protein is directed to two different compartments in the cell.

The mechanism for selecting a start codon in bacteria is different. Bacterial mRNAs have no 5′ caps to signal the ribosome where to begin searching for the start of translation. Instead, each bacterial mRNA contains a specific ribosome-binding site (called the Shine–Dalgarno sequence, named after its discoverers) that is located a few nucleotides upstream of the AUG at which translation is to begin. This nucleotide sequence, with the consensus 5′-AGGAGGU-3′, forms base pairs with the 16S rRNA of the small ribosomal subunit to position the initiating AUG codon in the ribosome. A set of translation initiation factors orchestrates this interaction, as well as the subsequent assembly of the large ribosomal subunit to complete the ribosome.

Unlike a eucaryotic ribosome, a bacterial ribosome can therefore readily assemble directly on a start codon that lies in the interior of an mRNA molecule, so long as a ribosome-binding site precedes it by several nucleotides. As a result, bacterial mRNAs are often *polycistronic*—that is, they encode several different proteins, each of which is translated from the same mRNA molecule (**Figure 6–73**). In contrast, a eucaryotic mRNA generally encodes only a single protein.

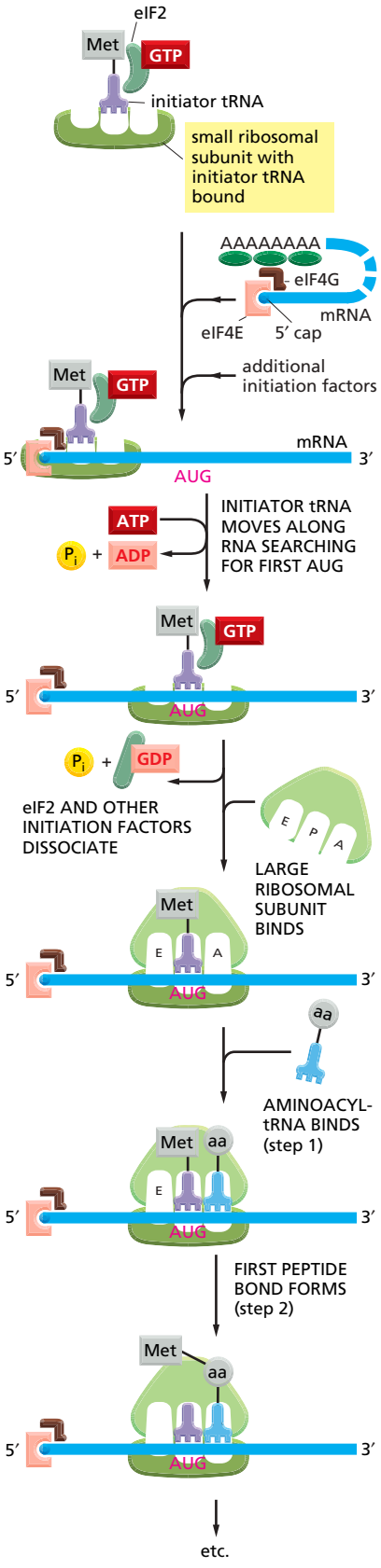


Figure 6–72 The initiation of protein synthesis in eucaryotes. Only three of the many translation initiation factors required for this process are shown. Efficient translation initiation also requires the poly-A tail of the mRNA bound by poly-A-binding proteins which, in turn, interact with eIF4G. In this way, the translation apparatus ascertains that both ends of the mRNA are intact before initiating protein synthesis (see Figure 6–40). Although only one GTP hydrolysis event is shown in the figure, a second is known to occur just before the large and small ribosomal subunits join.

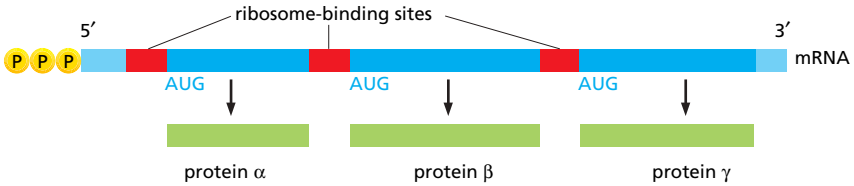


Figure 6–73 Structure of a typical bacterial mRNA molecule. Unlike eucaryotic ribosomes, which typically require a capped 5′ end, procaryotic ribosomes initiate transcription at ribosome-binding sites (Shine–Dalgarno sequences), which can be located anywhere along an mRNA molecule. This property of ribosomes permits bacteria to synthesize more than one type of protein from a single mRNA molecule.

Stop Codons Mark the End of Translation

The end of the protein-coding message is signaled by the presence of one of three *stop codons* (UAA, UAG, or UGA) (see Figure 6–50). These are not recognized by a tRNA and do not specify an amino acid, but instead signal to the ribosome to stop translation. Proteins known as *release factors* bind to any ribosome with a stop codon positioned in the A site, forcing the peptidyl transferase in the ribosome to catalyze the addition of a water molecule instead of an amino acid to the peptidyl-tRNA (**Figure 6–74**). This reaction frees the carboxyl end of the growing polypeptide chain from its attachment to a tRNA molecule, and since only this attachment normally holds the growing polypeptide to the ribosome, the completed protein chain is immediately released into the cytoplasm. The ribosome then releases the mRNA and separates into the large and small subunits, which can assemble on this or another mRNA molecule to begin a new round of protein synthesis.

Release factors are an example of *molecular mimicry*, whereby one type of macromolecule resembles the shape of a chemically unrelated molecule. In this case, the three-dimensional structure of release factors (made entirely of protein) resembles the shape and charge distribution of a tRNA molecule (**Figure 6–75**). This shape and charge mimicry helps them enter the A-site on the ribosome and cause translation termination.

During translation, the nascent polypeptide moves through a large, water-filled tunnel (approximately 10 nm × 1.5 nm) in the large subunit of the ribosome (see Figure 6–70C). The walls of this tunnel, made primarily of 23S rRNA, are a patchwork of tiny hydrophobic surfaces embedded in a more extensive hydrophilic surface. This structure is not complementary to any peptide, and thus provides a "Teflon" coating through which a polypeptide chain can easily slide. The dimensions of the tunnel suggest that nascent proteins are largely unstructured as they pass through the ribosome, although some α-helical regions of the protein can form before leaving the ribosome tunnel. As it leaves the ribosome, a newly synthesized protein must fold into its proper three-dimensional conformation to be useful to the cell, and later in this chapter we discuss how this folding occurs. First, however, we describe several additional aspects of the translation process itself.

Proteins Are Made on Polyribosomes

The synthesis of most protein molecules takes between 20 seconds and several minutes. During this very short period, however, it is usual for multiple initiations to take place on each mRNA molecule being translated. As soon as the preceding ribosome has translated enough of the nucleotide sequence to move out

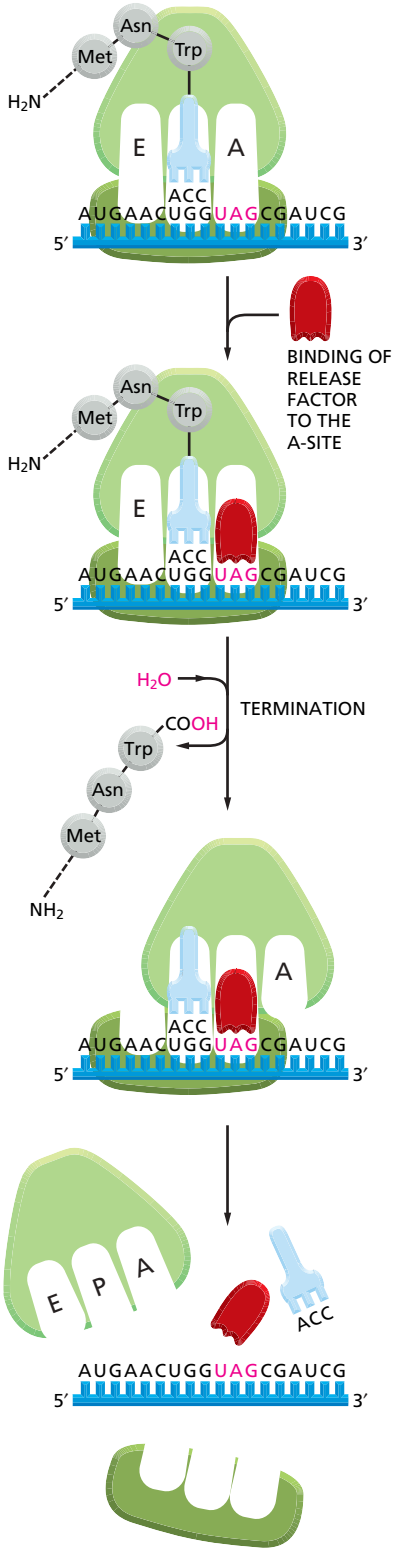


Figure 6–74 The final phase of protein synthesis. The binding of a release factor to an A-site bearing a stop codon terminates translation. The completed polypeptide is released and, in a series of reactions that requires additional proteins and GTP hydrolysis (not shown), the ribosome dissociates into its two separate subunits.

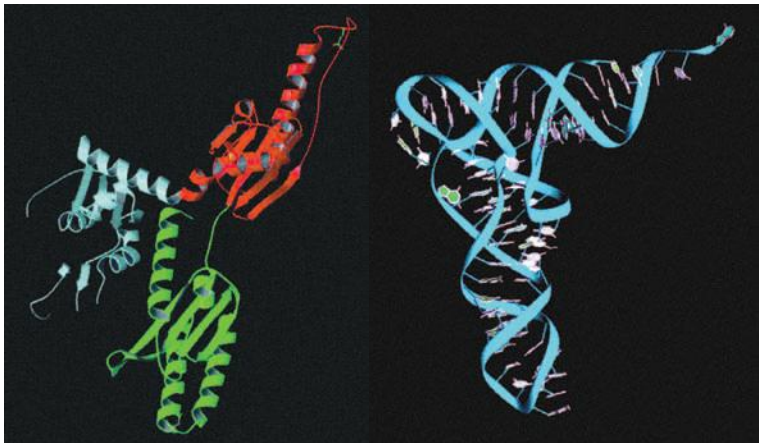


Figure 6–75 The structure of a human translation release factor (eRF1) and its resemblance to a tRNA molecule. The protein is on the *left* and the tRNA on the *right.* (From H. Song et al., *Cell* 100:311–321, 2000. With permission from Elsevier.)

of the way, the 5′ end of the mRNA is threaded into a new ribosome. The mRNA molecules being translated are therefore usually found in the form of *polyribosomes* (or *polysomes*): large cytoplasmic assemblies made up of several ribosomes spaced as close as 80 nucleotides apart along a single mRNA molecule (**Figure 6–76**). These multiple initiations allow the cell to make many more protein molecules in a given time than would be possible if each had to be completed before the next could start. <GAAG>

Both bacteria and eucaryotes use polysomes, and both employ additional strategies to speed up the overall rate of protein synthesis even further. Because bacterial mRNA does not need to be processed and is accessible to ribosomes while it is being made, ribosomes can attach to the free end of a bacterial mRNA molecule and start translating it even before the transcription of that RNA is complete, following closely behind the RNA polymerase as it moves along DNA. In eucaryotes, as we have seen, the 5′ and 3′ ends of the mRNA interact (see Figures 6–40 and 6–76A); therefore, as soon as a ribosome dissociates, its two subunits are in an optimal position to reinitiate translation on the same mRNA molecule.

There Are Minor Variations in the Standard Genetic Code

As discussed in Chapter 1, the genetic code (shown in Figure 6–50) applies to all three major branches of life, providing important evidence for the common

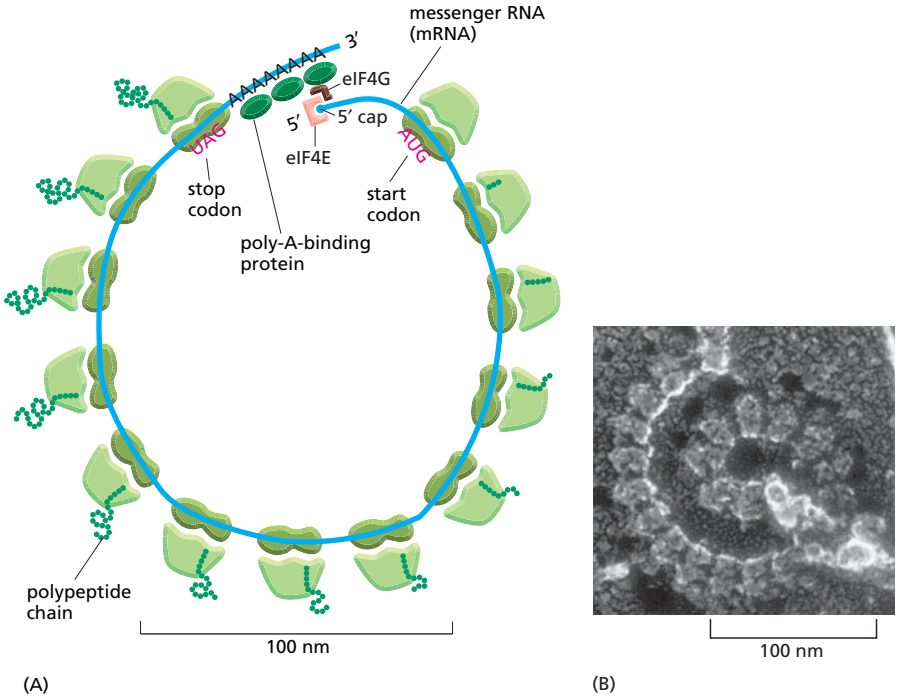


Figure 6–76 A polyribosome. (A) Schematic drawing showing how a series of ribosomes can simultaneously translate the same eucaryotic mRNA molecule. (B) Electron micrograph of a polyribosome from a eucaryotic cell. (B, courtesy of John Heuser.)

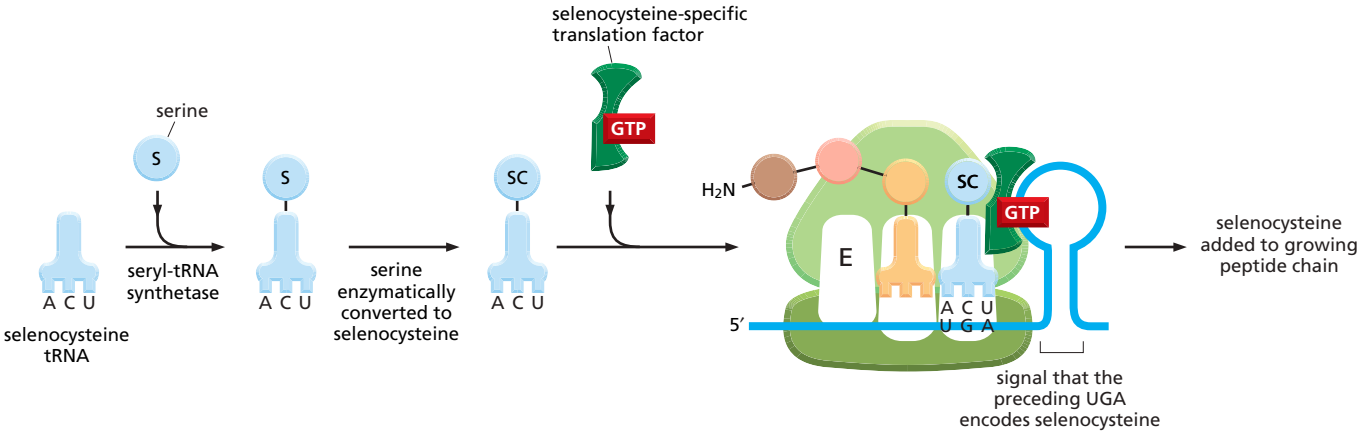


Figure 6–77 Incorporation of selenocysteine into a growing polypeptide chain. A specialized tRNA is charged with serine by the normal seryl-tRNA synthetase, and the serine is subsequently converted enzymatically to selenocysteine. A specific RNA structure in the mRNA (a stem and loop structure with a particular nucleotide sequence) signals that selenocysteine is to be inserted at the neighboring UGA codon. As indicated, this event requires the participation of a selenocysteine-specific translation factor.

ancestry of all life on Earth. Although rare, there are exceptions to this code. For example, *Candida albicans*, the most prevalent human fungal pathogen, translates the codon CUG as serine, whereas nearly all other organisms translate it as leucine. Mitochondria (which have their own genomes and encode much of their translational apparatus) often deviate from the standard code. For example, in mammalian mitochondria AUA is translated as methionine, whereas in the cytosol of the cell it is translated as isoleucine (see Table 14–3, p. 862). This type of deviation in the genetic code is "hardwired" into the organisms or the organelles in which it occurs.

A different type of variation, sometimes called *translation recoding*, occurs in many cells. In this case, other nucleotide sequence information present in an mRNA can change the meaning of the genetic code at a particular site in the mRNA molecule. The standard code allows cells to manufacture proteins using only 20 amino acids. However, bacteria, archaea, and eucaryotes have available to them a twenty-first amino acid that can be incorporated directly into a growing polypeptide chain through translation recoding. Selenocysteine, which is essential for the efficient function of a variety of enzymes, contains a selenium atom in place of the sulfur atom of cysteine. Selenocysteine is enzymatically produced from a serine attached to a special tRNA molecule that base-pairs with the UGA codon, a codon normally used to signal a translation stop. The mRNAs for proteins in which selenocysteine is to be inserted at a UGA codon carry an additional nucleotide sequence in the mRNA nearby that causes this recoding event (**Figure 6–77**).

Another form of recoding, *translational frameshifting*, allows more than one protein to be synthesized from a single mRNA. Retroviruses, members of a large group of eucaryotic-infecting pathogens, commonly use translational frameshifting to make both the capsid proteins *(Gag proteins)* and the viral reverse transcriptase and integrase *(Pol proteins)* from the same RNA transcript (see Figure 5–73). The virus needs many more copies of the Gag proteins than it does of the Pol proteins. This quantitative adjustment is achieved by encoding the *Pol* genes just after the *Gag* genes but in a different reading frame. Small amounts of the Pol gene products are made because, on occasion, an upstream translational frameshift allows the Gag protein stop codon to be bypassed. This frameshift occurs at a particular codon in the mRNA and requires a specific *recoding signal*, which seems to be a structural feature of the RNA sequence downstream of this site (**Figure 6–78**).

Inhibitors of Procaryotic Protein Synthesis Are Useful as Antibiotics

Many of the most effective antibiotics used in modern medicine are compounds made by fungi that inhibit bacterial protein synthesis. Fungi and bacteria compete for many of the same environmental niches, and millions of years of coevolution has resulted in fungi producing potent bacterial inhibitors. Some of these

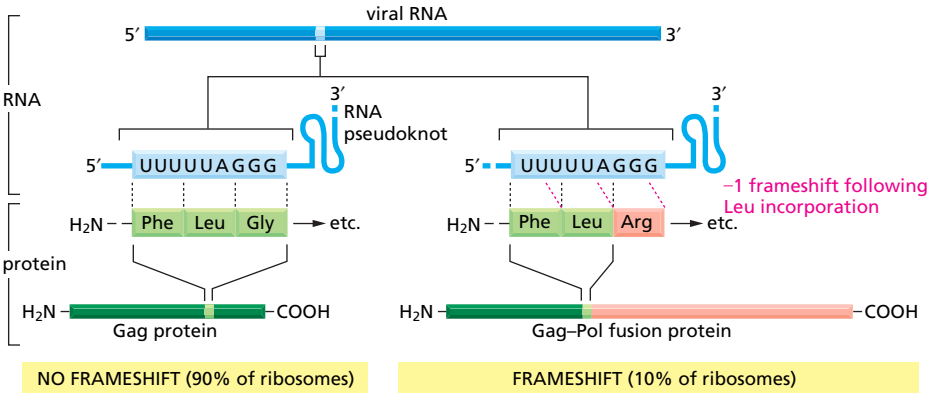


Figure 6–78 The translational frameshifting that produces the reverse transcriptase and integrase of a retrovirus. The viral reverse transcriptase and integrase are produced by proteolytic processing of a large protein (the Gag–Pol fusion protein) consisting of both the Gag and Pol amino acid sequences. Proteolytic processing of the more abundant Gag protein produces the viral capsid proteins. Both the Gag and the Gag–Pol fusion proteins start with identical mRNA, but whereas the Gag protein terminates at a stop codon downstream of the sequence shown, translation of the Gag–Pol fusion protein bypasses this stop codon, allowing the synthesis of the longer Gag–Pol fusion protein. The stop-codon-bypass is made possible by a controlled translational frameshift, as illustrated. Features in the local RNA structure (including the RNA loop shown) cause the $tRNA^{Leu}$ attached to the C-terminus of the growing polypeptide chain occasionally to slip backward by one nucleotide on the ribosome, so that it pairs with a UUU codon instead of the UUA codon that had initially specified its incorporation; the next codon (AGG) in the new reading frame specifies an arginine rather than a glycine. This controlled slippage is due in part to a *pseudoknot* that forms in the viral mRNA (see Figure 6–102). The sequence shown is from the human AIDS virus, HIV. (Adapted from T. Jacks et al., *Nature* 331:280–283, 1988. With permission from Macmillan Publishers Ltd.)

drugs exploit the structural and functional differences between bacterial and eucaryotic ribosomes so as to interfere preferentially with the function of bacterial ribosomes. Thus humans can take high dosages of some of these compounds without undue toxicity. Many antibiotics lodge in pockets in the ribosomal RNAs and simply interfere with the smooth operation of the ribosome (**Figure 6–79**). **Table 6–4** lists some of the more common antibiotics of this kind along with several other inhibitors of protein synthesis, some of which act on eucaryotic cells and therefore cannot be used as antibiotics.

Because they block specific steps in the processes that lead from DNA to protein, many of the compounds listed in Table 6–4 are useful for cell biological studies. Among the most commonly used drugs in such investigations are *chloramphenicol*, *cycloheximide*, and *puromycin*, all of which specifically inhibit protein synthesis. In a eucaryotic cell, for example, chloramphenicol inhibits protein synthesis on ribosomes only in mitochondria (and in chloroplasts in plants), presumably reflecting the procaryotic origins of these organelles (discussed in Chapter 14). Cycloheximide, in contrast, affects only ribosomes in the cytosol. Puromycin is especially interesting because it is a structural analog of a tRNA molecule linked to an amino acid and is therefore another example of molecular mimicry; the ribosome mistakes it for an authentic amino acid and covalently incorporates it at the C-terminus of the growing peptide chain, thereby causing the premature termination and release of the polypeptide. As might be expected, puromycin inhibits protein synthesis in both procaryotes and eucaryotes.

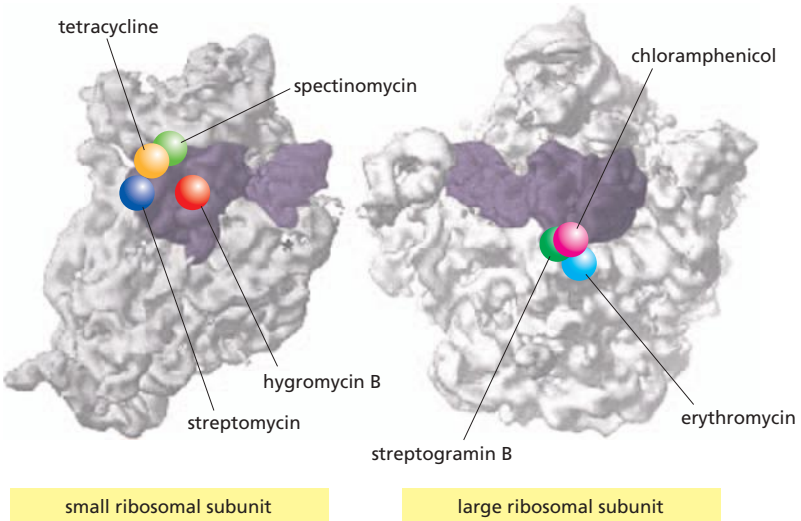


Figure 6–79 Binding sites for antibiotics on the bacterial ribosome. The small *(left)* and large *(right)* subunits of the ribosome are arranged as though the ribosome has been opened like a book; the bound tRNA molecules are shown in *purple* (see Figure 6–64). Most of the antibiotics shown bind directly to pockets formed by the ribosomal RNA molecules. Hygromycin B induces errors in translation, spectinomycin blocks the translocation of the peptidyl-tRNA from the A-site to the P-site, and streptogramin B prevents elongation of nascent peptides. Table 6–4 lists the inhibitory mechanisms of the other antibiotics shown in the figure. (Adapted from J. Poehlsgaard and S. Douthwaite, *Nat. Rev. Microbiol.* 3:870–881, 2005. With permission from Macmillan Publishers Ltd.)

Table 6–4 Inhibitors of Protein or RNA Synthesis

INHIBITOR	SPECIFIC EFFECT
Acting only on bacteria	
Tetracycline	blocks binding of aminoacyl-tRNA to A-site of ribosome
Streptomycin	prevents the transition from translation initiation to chain elongation and also causes miscoding
Chloramphenicol	blocks the peptidyl transferase reaction on ribosomes (step 2 in Figure 6–66)
Erythromycin	binds in the exit channel of the ribosome and thereby inhibits elongation of the peptide chain
Rifamycin	blocks initiation of RNA chains by binding to RNA polymerase (prevents RNA synthesis)
Acting on bacteria and eucaryotes	
Puromycin	causes the premature release of nascent polypeptide chains by its addition to the growing chain end
Actinomycin D	binds to DNA and blocks the movement of RNA polymerase (prevents RNA synthesis)
Acting on eucaryotes but not bacteria	
Cycloheximide	blocks the translocation reaction on ribosomes (step 3 in Figure 6–66)
Anisomycin	blocks the peptidyl transferase reaction on ribosomes (step 2 in Figure 6–66)
α-Amanitin	blocks mRNA synthesis by binding preferentially to RNA polymerase II

The ribosomes of eucaryotic mitochondria (and chloroplasts) often resemble those of bacteria in their sensitivity to inhibitors. Therefore, some of these antibiotics can have a deleterious effect on human mitochondria.

Accuracy in Translation Requires the Expenditure of Free Energy

Translation by the ribosome is a compromise between the opposing constraints of accuracy and speed. We have seen, for example, that the accuracy of translation (1 mistake per 10^4 amino acids joined) requires time delays each time a new amino acid is added to a growing polypeptide chain, producing an overall speed of translation of 20 amino acids incorporated per second in bacteria. Mutant bacteria with a specific alteration in the small ribosomal subunit have longer delays and translate mRNA into protein with an accuracy considerably higher than this; however, protein synthesis is so slow in these mutants that the bacteria are barely able to survive.

We have also seen that attaining the observed accuracy of protein synthesis requires the expenditure of a great deal of free energy; this is expected, since, as discussed in Chapter 2, there is a price to be paid for any increase in order in the cell. In most cells, protein synthesis consumes more energy than any other biosynthetic process. At least four high-energy phosphate bonds are split to make each new peptide bond: two are consumed in charging a tRNA molecule with an amino acid (see Figure 6–56), and two more drive steps in the cycle of reactions occurring on the ribosome during synthesis itself (see Figure 6–67). In addition, extra energy is consumed each time that an incorrect amino acid linkage is hydrolyzed by a tRNA synthetase (see Figure 6–59) and each time that an incorrect tRNA enters the ribosome, triggers GTP hydrolysis, and is rejected (see Figure 6–67). To be effective, these proofreading mechanisms must also allow an appreciable fraction of correct interactions to be removed; for this reason, proofreading is even more costly in energy than it might seem.

Quality Control Mechanisms Act to Prevent Translation of Damaged mRNAs

In eucaryotes, mRNA production involves both transcription and a series of elaborate RNA-processing steps; these take place in the nucleus, segregated from ribosomes, and only when the processing is complete are the mRNAs transported to the cytoplasm to be translated (see Figure 6–40). However, this scheme is not foolproof, and some incorrectly processed mRNAs are inadvertently sent to the cytoplasm. In addition, mRNAs that were flawless when they left the nucleus can become broken or otherwise damaged in the cytosol. The danger of translating damaged or incompletely processed mRNAs (which would produce truncated or otherwise aberrant proteins) is apparently so great that the cell has several backup measures to prevent this from happening.

To avoid translating broken mRNAs, the 5′ cap and the poly-A tail are both recognized by the translation-initiation machinery before translation begins (see Figure 6–72). To help ensure that mRNAs are properly spliced before they are translated, the exon junction complex (EJC), which is deposited on the mRNA following splicing (see Figure 6–40), stimulates the subsequent translation of the mRNA.

But the most powerful mRNA surveillance system, called **nonsense-mediated mRNA decay**, eliminates defective mRNAs before they can be efficiently translated into protein. This mechanism is brought into play when the cell determines that an mRNA molecule has a nonsense (stop) codon (UAA, UAG, or UGA) in the "wrong" place—a situation likely to arise in an mRNA molecule that has been improperly spliced. Aberrant splicing will usually result in the random introduction of a nonsense codon into the reading frame of the mRNA, especially in organisms, such as humans, that have a large average intron size (see Figure 6–32B).

This surveillance mechanism begins as an mRNA molecule is being transported from the nucleus to the cytosol. As its 5′ end emerges from the nuclear pore, the mRNA is met by a ribosome, which begins to translate it. As translation proceeds, the exon junction complexes (EJC) bound to the mRNA at each splice-site are apparently displaced by the moving ribosome. The normal stop codon will be within the last exon, so by the time the ribosome reaches it and stalls, no more EJCs should be bound to the mRNA. If this is the case, the mRNA "passes inspection" and is released to the cytosol where it can be translated in earnest (**Figure 6–80**). However, if the ribosome reaches a premature stop codon and stalls, it senses that EJCs remain and the bound mRNA molecule is rapidly degraded. In this way, the first round of translation allows the cell to test the fitness of each mRNA molecule as it exits the nucleus.

Nonsense-mediated decay may have been especially important in evolution, allowing eucaryotic cells to more easily explore new genes formed by DNA rearrangements, mutations, or alternative patterns of splicing—by selecting only those mRNAs for translation that can produce a full-length protein. Nonsense-mediated decay is also important in cells of the developing immune system, where the extensive DNA rearrangements that occur (see Figure 25–36) often generate premature termination codons. The surveillance system degrades the mRNAs produced from such rearranged genes, thereby avoiding the potential toxic effects of truncated proteins.

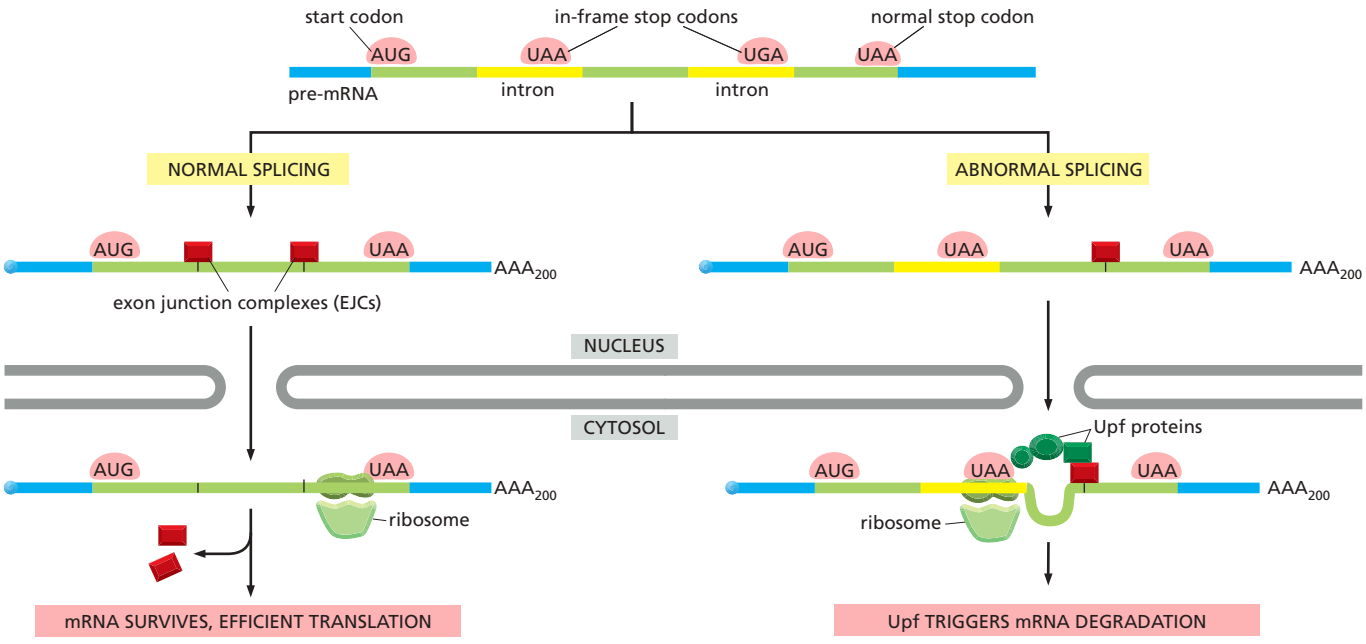


Figure 6–80 Nonsense-mediated mRNA decay. As shown on the right, the failure to correctly splice a pre-mRNA often introduces a premature stop codon into the reading frame for the protein. The introduction of such an "in-frame" stop codon is particularly likely to occur in mammals, where the introns tend to be very long. When translated, these abnormal mRNAs produce aberrant proteins, which could damage the cell. However, as shown at the bottom right of the figure, these abnormal RNAs are destroyed by the nonsense-mediated decay mechanism. According to one model, an mRNA molecule, bearing exon junction complexes (EJCs) to mark successfully completed splices, is first met by a ribosome that performs a "test" round of translation. As the mRNA passes through the tight channel of the ribosome, the EJCs are stripped off, and successful mRNAs are released to undergo multiple rounds of translation *(left side)*. However, if an in-frame stop codon is encountered before the final exon junction complex is reached (*right side*), the mRNA undergoes nonsense-mediated decay, which is triggered by the Upf proteins (*green*) that bind to each EJC. Note that, to trigger nonsense-mediated decay, the premature stop codon must be in the same reading frame as that of the normal protein. (Adapted from J. Lykke-Andersen et al., *Cell* 103:1121–1131, 2000. With permission from Elsevier.)

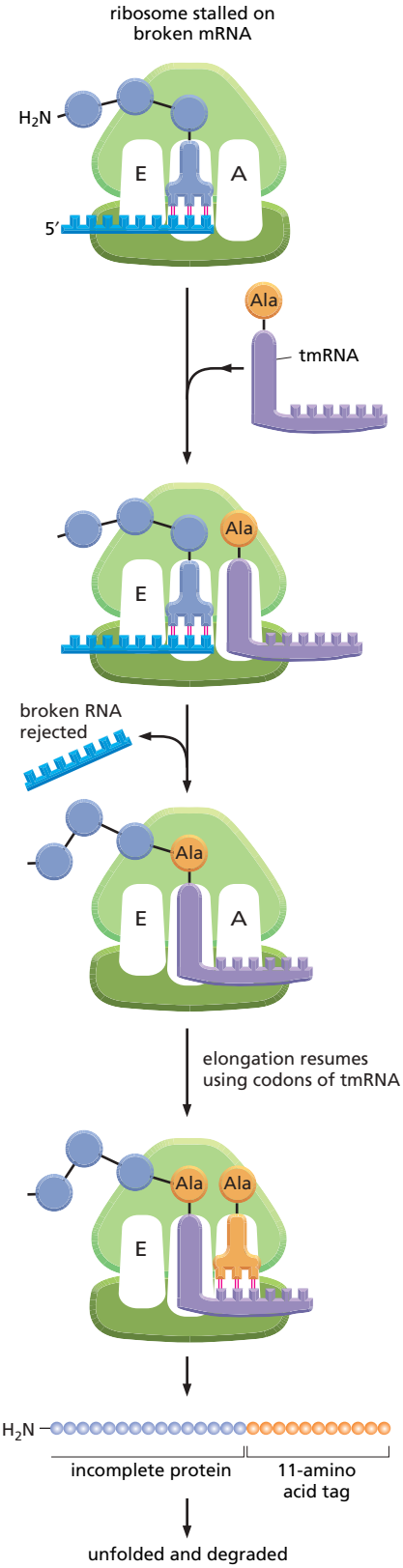


Figure 6–81 The rescue of a bacterial ribosome stalled on an incomplete mRNA molecule. The tmRNA shown is a 363-nucleotide RNA with both tRNA and mRNA functions, hence its name. It carries an alanine and can enter the vacant A-site of a stalled ribosome to add this alanine to a polypeptide chain, mimicking a tRNA although no codon is present to guide it. The ribosome then translates 10 codons from the tmRNA, completing an 11 amino acid tag on the protein. Proteases recognize this tag and degrade the entire protein. Although the example shown in the figure is from bacteria, eucaryotes can employ a similar strategy.

Finally, the nonsense-mediated surveillance pathway plays an important role in mitigating the symptoms of many inherited human diseases. As we have seen, inherited diseases are usually caused by mutations that spoil the function of a key protein, such as hemoglobin or one of the blood clotting factors. Approximately one-third of all genetic disorders in humans result from nonsense mutations or mutations (such as frameshift mutations or splice-site mutations) that place nonsense mutations into the gene's reading frame. In individuals that carry one mutant and one functional gene, nonsense-mediated decay eliminates the aberrant mRNA and thereby prevents a potentially toxic protein from being made. Without this safeguard, individuals with one functional and one mutant "disease gene" would likely suffer much more severe symptoms.

We saw earlier in this chapter that bacteria lack the elaborate mRNA processing found in eucaryotes and that translation often begins before the synthesis of the RNA molecule is completed. Yet bacteria also have quality control mechanisms to deal with incompletely synthesized and broken mRNAs. When the bacterial ribosome translates to the end of an incomplete RNA it stalls and does not release the RNA. Rescue comes in the form of a special RNA (called tmRNA), which enters the A-site of the ribosome and is itself translated, releasing the ribosome. The special 11 amino acid tag thus added to the C-terminus of the truncated protein signals to proteases that the entire protein is to be degraded (**Figure 6–81**).

Some Proteins Begin to Fold While Still Being Synthesized

The process of gene expression is not over when the genetic code has been used to create the sequence of amino acids that constitutes a protein. To be useful to the cell, this new polypeptide chain must fold up into its unique three-dimensional conformation, bind any small-molecule cofactors required for its activity, be appropriately modified by protein kinases or other protein-modifying enzymes, and assemble correctly with the other protein subunits with which it functions (**Figure 6–82**).

The information needed for all of the steps listed above is ultimately contained in the sequence of linked amino acids that the ribosome produces when it translates an mRNA molecule into a polypeptide chain. As discussed in Chapter 3, when a protein folds into a compact structure, it buries most of its hydrophobic residues in an interior core. In addition, large numbers of noncovalent interactions form between various parts of the molecule. It is the sum of all of these energetically favorable arrangements that determines the final folding pattern of the polypeptide chain—as the conformation of lowest free energy (see p. 130).

Through many millions of years of evolution, the amino acid sequence of each protein has been selected not only for the conformation that it adopts but also for an ability to fold rapidly. For some proteins, this folding begins immediately, as the protein spins out of the ribosome, starting from the N-terminal end. In these cases, as each protein domain emerges from the ribosome, within a few seconds it forms a compact structure that contains most of the final secondary features (α helices and β sheets) aligned in roughly the right conformation (**Figure 6–83**). For many protein domains, this unusually dynamic and flexible state called a *molten globule*, is the starting point for a relatively slow process in which many side-chain adjustments occur that eventually form the correct tertiary

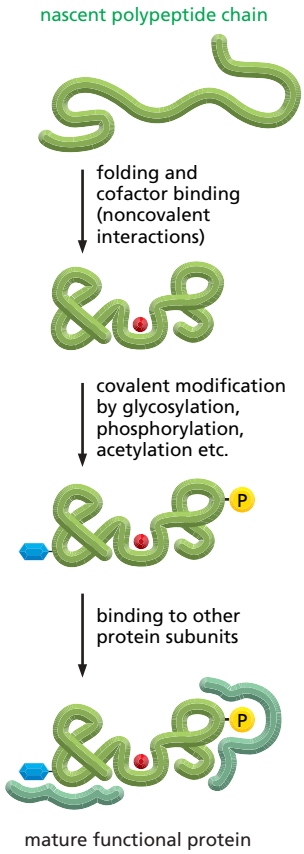


Figure 6–82 Steps in the creation of a functional protein. As indicated, translation of an mRNA sequence into an amino acid sequence on the ribosome is not the end of the process of forming a protein. To function, the completed polypeptide chain must fold correctly into its three-dimensional conformation, bind any cofactors required, and assemble with its partner protein chains (if any). Noncovalent bond formation drives these changes. As indicated, many proteins also require covalent modifications of selected amino acids. Although the most frequent modifications are protein glycosylation and protein phosphorylation, more than 100 different types of covalent modifications are known (see, for example, Figure 3–81).

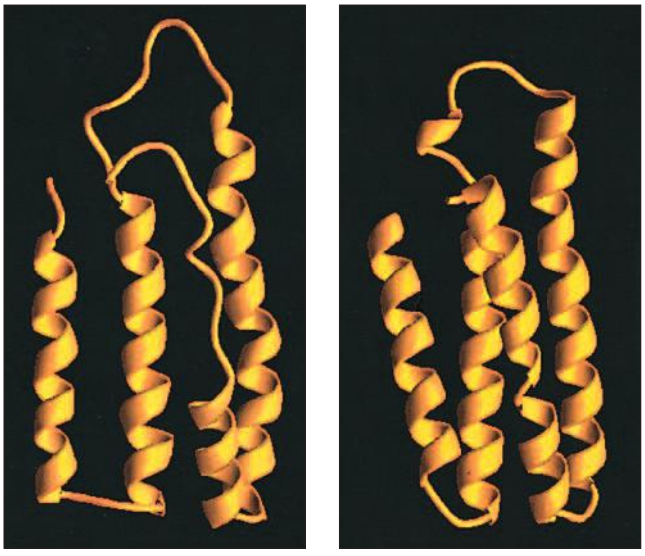
structure. It takes several minutes to synthesize a protein of average size, and for some proteins much of the folding process is complete by the time the ribosome releases the C-terminal end of a protein (**Figure 6–84**).

Molecular Chaperones Help Guide the Folding of Most Proteins

Most proteins probably do not begin to fold during their synthesis. Instead, they are met at the ribosome by a special class of proteins called **molecular chaperones**. Molecular chaperones are useful for cells because there are many different paths that can be taken to convert an unfolded or partially folded protein to its final compact conformation. For many proteins, some of the intermediates formed along the way would aggregate and be left as off-pathway dead ends without the intervention of a chaperone (**Figure 6–85**).

Many molecular chaperones are called *heat-shock proteins* (designated *Hsp*), because they are synthesized in dramatically increased amounts after a brief exposure of cells to an elevated temperature (for example, 42°C for cells that normally live at 37°C). This reflects the operation of a feedback system that responds to an increase in misfolded proteins (such as those produced by elevated temperatures) by boosting the synthesis of the chaperones that help these proteins refold.

There are several major families of eucaryotic molecular chaperones, including the Hsp60 and Hsp70 proteins. Different family members function in different organelles. Thus, as discussed in Chapter 12, mitochondria contain their own Hsp60 and Hsp70 molecules that are distinct from those that function in the cytosol; and a special Hsp70 (called *BIP*) helps to fold proteins in the endoplasmic reticulum.



(A) (B)

Figure 6–83 The structure of a molten globule. (A) A molten globule form of cytochrome b_{562} is more open and less highly ordered than the final folded form of the protein, shown in (B). Note that the molten globule contains most of the secondary structure of the final form, although the ends of the α helices are unravelled and one of the helices is only partly formed. (Courtesy of Joshua Wand, from Y. Feng et al., *Nat. Struct. Biol.* 1:30–35, 1994. With permission from Macmillan Publishers Ltd.)

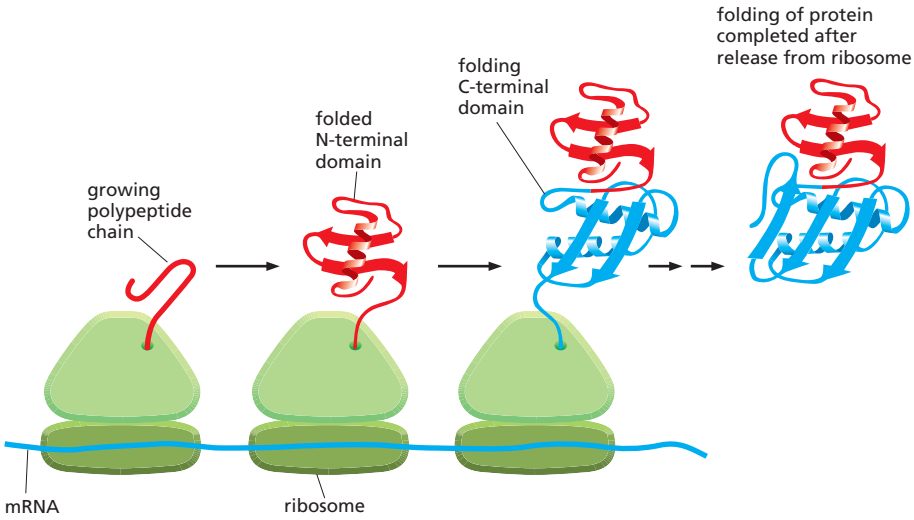


Figure 6–84 Co-translational protein folding. A growing polypeptide chain is shown acquiring its secondary and tertiary structure as it emerges from a ribosome. The N-terminal domain folds first, while the C-terminal domain is still being synthesized. This protein has not achieved its final conformation at the time it is released from the ribosome. (Modified from A.N. Federov and T.O. Baldwin, *J. Biol. Chem.* 272:32715–32718, 1997.)

The Hsp60 and Hsp70 proteins each work with their own small set of associated proteins when they help other proteins to fold. Hsps share an affinity for the exposed hydrophobic patches on incompletely folded proteins, and they hydrolyze ATP, often binding and releasing their protein substrate with each cycle of ATP hydrolysis. In other respects, the two types of Hsp proteins function differently. The Hsp70 machinery acts early in the life of many proteins, binding to a string of about seven hydrophobic amino acids before the protein leaves the ribosome (**Figure 6–86**). In contrast, Hsp60-like proteins form a large barrel-shaped structure that acts after a protein has been fully synthesized. This type of chaperone, sometimes called a *chaperonin*, forms an "isolation chamber" into which misfolded proteins are fed, preventing their aggregation and providing them with a favorable environment in which to attempt to refold (**Figure 6–87**).

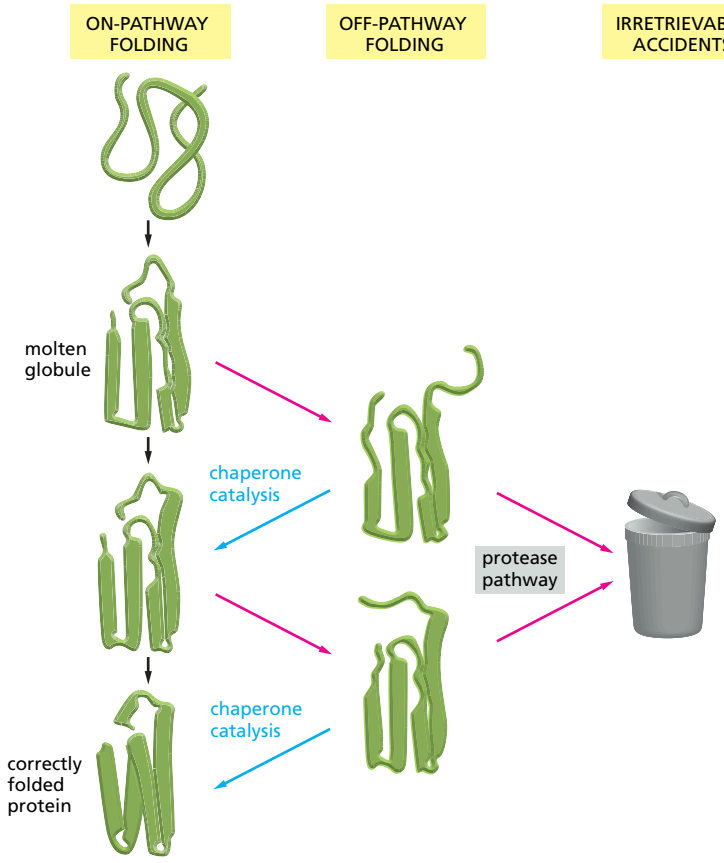


Figure 6–85 A current view of protein folding. Each domain of a newly synthesized protein rapidly attains a "molten globule" state. Subsequent folding occurs more slowly and by multiple pathways, often involving the help of a molecular chaperone. Some molecules may still fail to fold correctly; as explained in the text, specific proteases recognize and degrade these molecules.

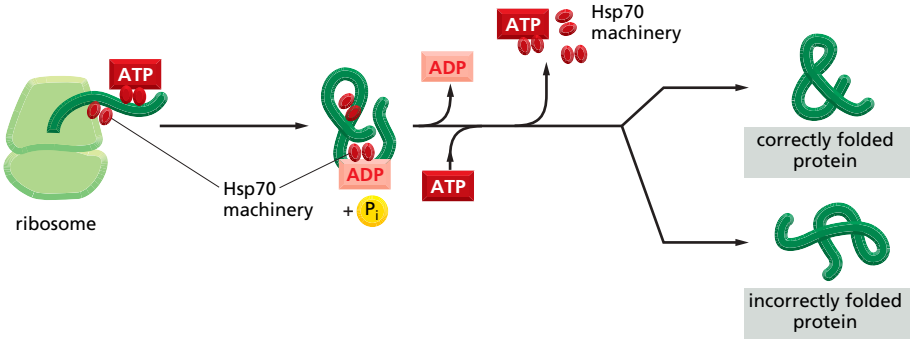


Figure 6–86 The Hsp70 family of molecular chaperones. These proteins act early, recognizing a small stretch of hydrophobic amino acids on a protein's surface. Aided by a set of smaller Hsp40 proteins (not shown), ATP-bound Hsp70 molecules grasp their target protein and then hydrolyze ATP to ADP, undergoing conformational changes that cause the Hsp70 molecules to associate even more tightly with the target. After the Hsp40 dissociates, the rapid rebinding of ATP induces the dissociation of the Hsp70 protein after ADP release. In reality, repeated cycles of Hsp protein binding and release help the target protein to refold, as schematically illustrated in Figure 6–85.

The chaperones shown in Figures 6–86 and 6–87 often use many cycles of ATP hydrolysis to fold a single polypeptide chain correctly. Although some of this energy expenditure is used to perform mechanical work, probably much more is expended to ensure that protein folding is accurate. Just as we saw for transcription, splicing, and translation, the expenditure of free energy can be used by cells to improve the accuracy of a biological process. In the case of protein folding, ATP hydrolysis allows chaperones to recognize a wide variety of misfolded structures, to halt any further misfolding and to recommence folding of a protein in an orderly way.

Although our discussion focuses on only two types of chaperones, the cell has a variety of others. The enormous diversity of proteins in cells presumably requires a wide range of chaperones with versatile surveillance and correction capabilities.

Exposed Hydrophobic Regions Provide Critical Signals for Protein Quality Control

If radioactive amino acids are added to cells for a brief period, the newly synthesized proteins can be followed as they mature into their final functional form.

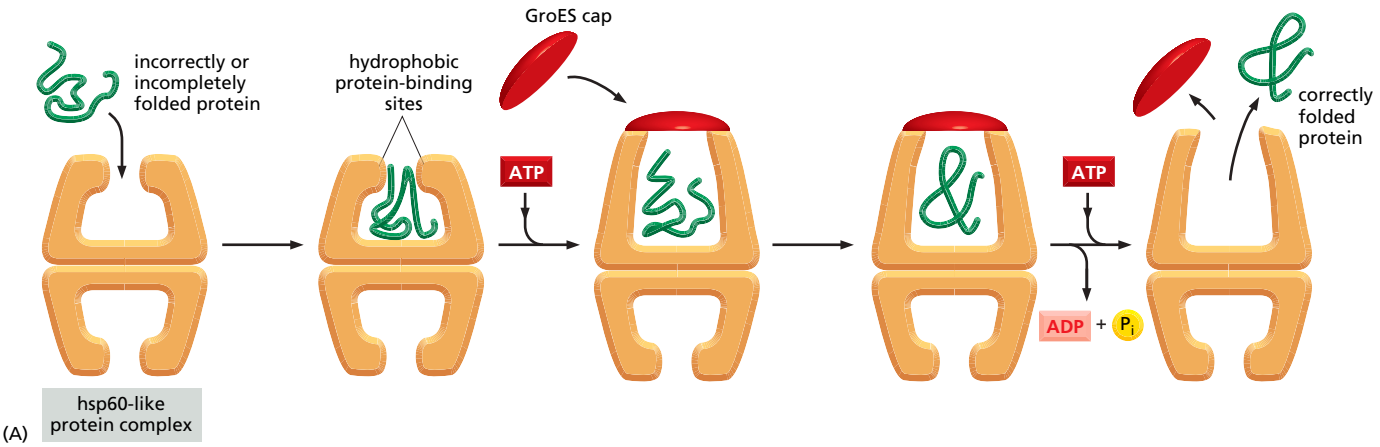
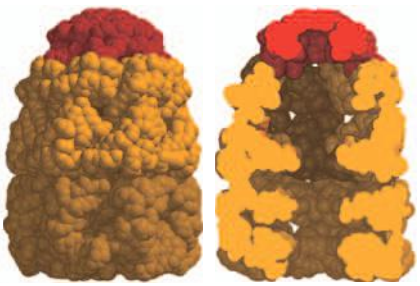


Figure 6–87 The structure and function of the Hsp60 family of molecular chaperones. (A) The catalysis of protein refolding. A misfolded protein is initially captured by hydrophobic interactions along one rim of the barrel. The subsequent binding of ATP plus a protein cap increases the diameter of the barrel rim, which may transiently stretch (partly unfold) the client protein. This also confines the protein in an enclosed space, where it has a new opportunity to fold. After about 15 seconds, ATP hydrolysis occurs, weakening the complex. Subsequent binding of another ATP molecule ejects the protein, whether folded or not, and the cycle repeats. This type of molecular chaperone is also known as a chaperonin; it is designated as Hsp60 in mitochondria, TCP1 in the cytosol of vertebrate cells, and GroEL in bacteria. As indicated, only half of the symmetrical barrel operates on a client protein at any one time. (B) The structure of GroEL bound to its GroES cap, as determined by X-ray crystallography. On the *left* is shown the outside of the barrel-like structure and on the *right* a cross section through its center. (B, adapted from B. Bukau and A.L. Horwich, *Cell* 92:351–366, 1998. With permission from Elsevier.)



This type of experiment demonstrates that the Hsp70 proteins act first, beginning when a protein is still being synthesized on a ribosome, and the Hsp60-like proteins act only later to help fold completed proteins. But how does the cell distinguish misfolded proteins, which require additional rounds of ATP-catalyzed refolding, from those with correct structures?

Before answering, we need to pause to consider the post-translational fate of proteins more broadly. Usually, if a protein has a sizable exposed patch of hydrophobic amino acids on its surface, it is abnormal: it has either failed to fold correctly after leaving the ribosome, suffered an accident that partly unfolded it at a later time, or failed to find its normal partner subunit in a larger protein complex. Such a protein is not merely useless to the cell, it can be dangerous. Many proteins with an abnormally exposed hydrophobic region can form large aggregates in the cell. We shall see that, in rare cases, such aggregates do form and cause severe human diseases. Normally, however, powerful protein quality control mechanisms prevent such disasters.

Given this background, it is not surprising that cells have evolved elaborate mechanisms that recognize the hydrophobic patches on proteins and minimize the damage they cause. Two of these mechanisms depend on the molecular chaperones just discussed, which bind to the patch and attempt to repair the defective protein by giving it another chance to fold. At the same time, by covering the hydrophobic patches, these chaperones transiently prevent protein aggregation. Proteins that very rapidly fold correctly on their own do not display such patches and the chaperones bypass them.

Figure 6–88 outlines all of the quality control choices that a cell makes for a difficult-to-fold, newly synthesized protein. As indicated, when attempts to refold a protein fail, a third mechanism is called into play that completely destroys the protein by proteolysis. The proteolytic pathway begins with the recognition of an abnormal hydrophobic patch on a protein's surface, and it ends with the delivery of the entire protein to a protein destruction machine, a complex protease known as the *proteasome*. As described next, this process depends on an elaborate protein-marking system that also carries out other central functions in the cell by destroying selected normal proteins.

The Proteasome Is a Compartmentalized Protease with Sequestered Active Sites

The proteolytic machinery and the chaperones compete with one another to reorganize a misfolded protein. If a newly synthesized protein folds rapidly, at most only a small fraction of it is degraded. In contrast, a slowly folding protein is vulnerable to the proteolytic machinery for a longer time, and many more of its molecules are destroyed before the remainder attain the proper folded state. Due to mutations or to errors in transcription, RNA splicing, and translation, some proteins never fold properly. It is particularly important that the cell destroy these potentially harmful proteins.

The apparatus that deliberately destroys aberrant proteins is the **proteasome**, an abundant ATP-dependent protease that constitutes nearly 1% of cell protein. Present in many copies dispersed throughout the cytosol and the nucleus, the proteasome also destroys aberrant proteins of the endoplasmic

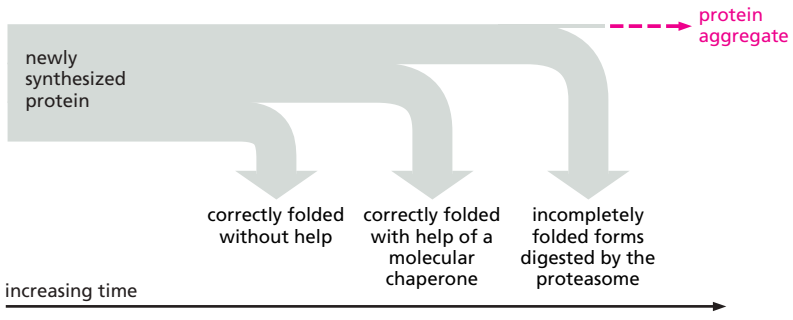


Figure 6–88 The processes that monitor protein quality following protein synthesis. A newly synthesized protein sometimes folds correctly and assembles on its own with its partner proteins, in which case the quality control mechanisms leave it alone. Incompletely folded proteins are helped to refold by molecular chaperones: first by a family of Hsp70 proteins, and then in some cases, by Hsp60-like proteins. For both types of chaperones, the client proteins are recognized by an abnormally exposed patch of hydrophobic amino acids on their surface. These "protein-rescue" processes compete with another mechanism that, upon recognizing an abnormally exposed patch, marks the protein for destruction by the proteasome. The combined activity of all of these processes is needed to prevent massive protein aggregation in a cell, which can occur when many hydrophobic regions on proteins clump together nonspecifically.

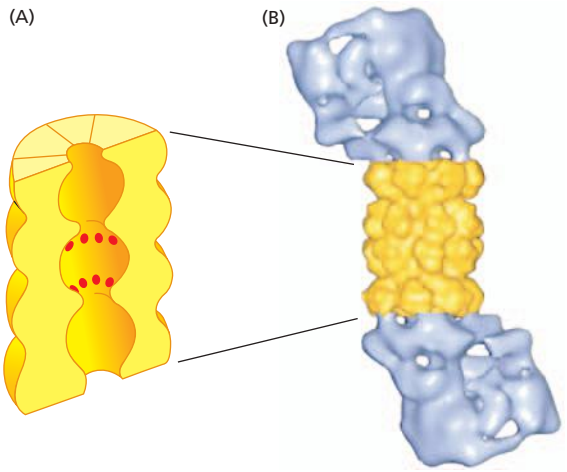


Figure 6–89 The proteasome. (A) A cutaway view of the structure of the central 20S cylinder, as determined by x-ray crystallography, with the active sites of the proteases indicated by *red dots*. (B) The entire proteasome, in which the central cylinder *(yellow)* is supplemented by a 19S cap *(blue)* at each end. The cap structure has been determined by computer processing of electron microscope images. The complex cap (also called the regulatory particle) selectively binds proteins that have been marked by ubiquitin for destruction; it then uses ATP hydrolysis to unfold their polypeptide chains and feed them through a narrow channel (see Figure 6–91) into the inner chamber of the 20S cylinder for digestion to short peptides. (B, from W. Baumeister et al., *Cell* 92:367–380, 1998. With permission from Elsevier.)

reticulum (ER). An ER-based surveillance system detects proteins that fail either to fold or to be assembled properly after they enter the ER, and *retrotranslocates* them back to the cytosol for degradation (discussed in Chapter 12).

Each proteasome consists of a central hollow cylinder (the 20S core proteasome) formed from multiple protein subunits that assemble as a quasi-cylindrical stack of four heptameric rings (**Figure 6–89**). Some of the subunits are distinct proteases whose active sites face the cylinder's inner chamber. The design prevents these highly efficient proteases from running rampant through the cell. Each end of the cylinder is normally associated with a large protein complex (the 19S cap), which contains a six-subunit protein ring, through which target proteins are threaded into the proteasome core where they are degraded (**Figure 6–90**). The threading reaction, driven by ATP hydrolysis, unfolds the target proteins as they move through the cap, exposing them to the proteases lining the proteasome core (**Figure 6–91**). The proteins that make up the ring structure in the proteasome cap belong to a large class of protein "unfoldases" known as *AAA proteins*. Many of them function as hexamers, and it is possible that they share mechanistic features with the ATP-dependent unwinding of DNA by DNA helicases (see Figure 5–15).

A crucial property of the proteasome, and one reason for the complexity of its design, is the *processivity* of its mechanism: in contrast to a "simple" protease that cleaves a substrate's polypeptide chain just once before dissociating, the proteasome keeps the entire substrate bound until all of it is converted into short peptides.

The 19S caps also act as regulated "gates" at the entrances to the inner proteolytic chamber, and they are responsible for binding a targeted protein substrate to the proteasome. With a few exceptions, the proteasomes act on proteins that have been specifically marked for destruction by the covalent attachment of a recognition tag formed from a small protein called *ubiquitin* (**Figure 6–92**A). Ubiquitin exists in cells either free or covalently linked to

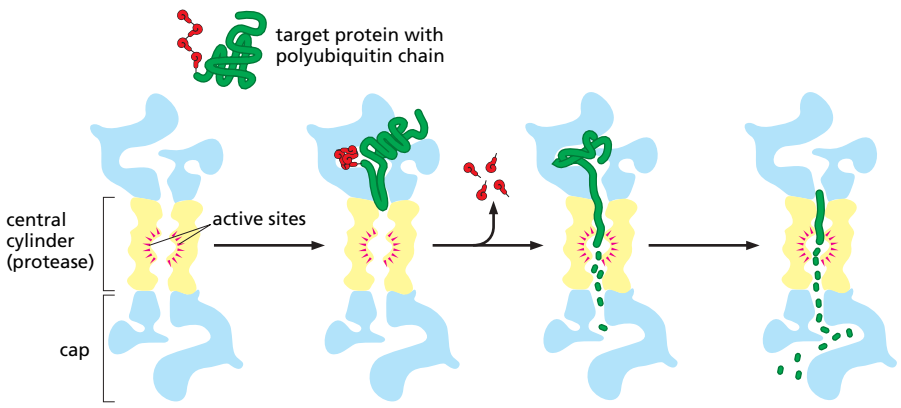


Figure 6–90 Processive protein digestion by the proteasome. The proteasome cap recognizes a substrate protein, in this case marked by a polyubiquitin chain (see Figure 6–92), and subsequently translocates it into the proteasome core, where it is digested. At an early stage, the ubiquitin is cleaved from the substrate protein and is recycled. Translocation into the core of the proteasome is mediated by a ring of ATP-dependent proteins that unfold the substrate protein as it is threaded through the ring and into the proteasome core (see Figure 6–91). (From S. Prakash and A. Matouschek, *Trends Biochem. Sci.* 29:593–600, 2004. With permission from Elsevier.)

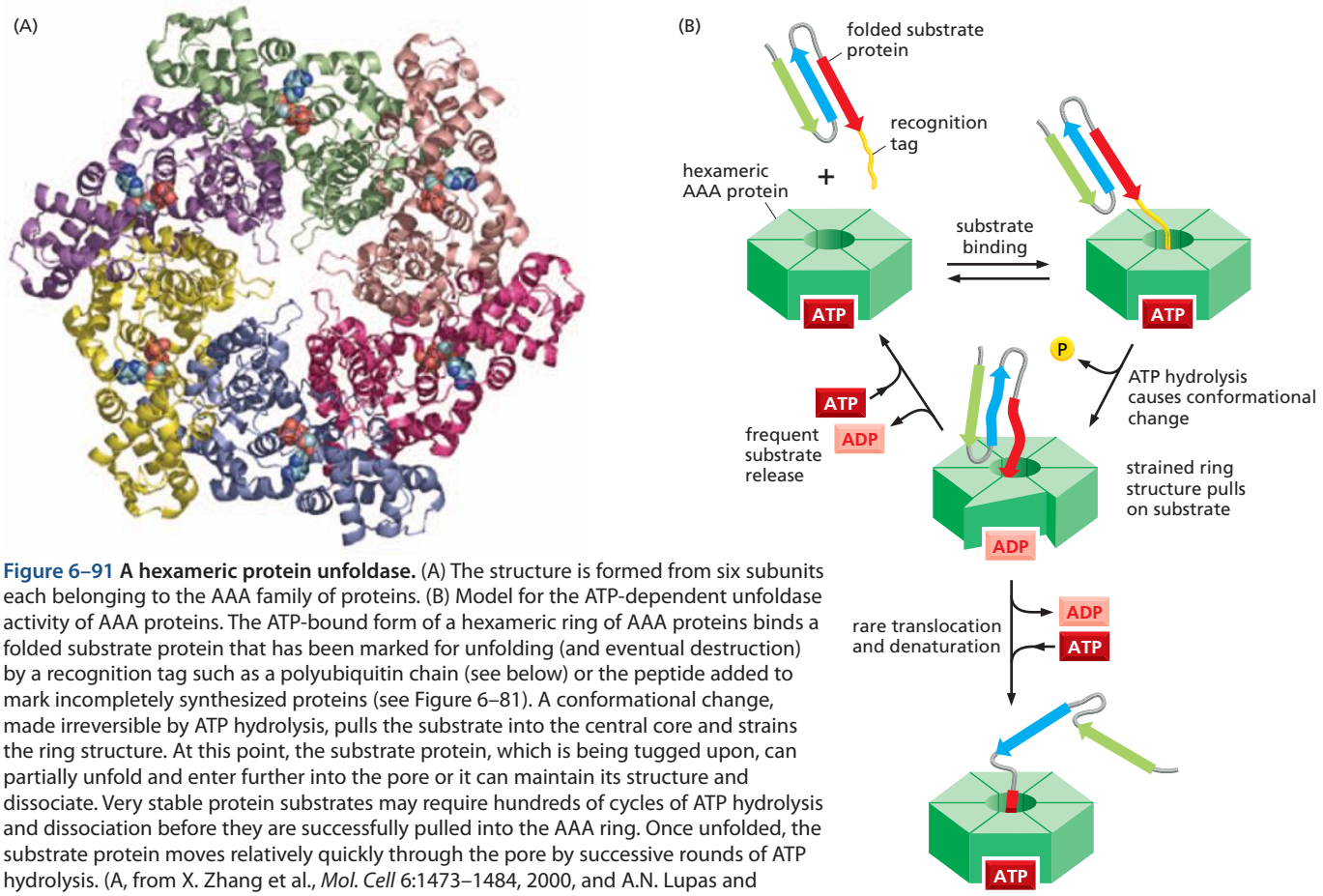


Figure 6–91 A hexameric protein unfoldase. (A) The structure is formed from six subunits each belonging to the AAA family of proteins. (B) Model for the ATP-dependent unfoldase activity of AAA proteins. The ATP-bound form of a hexameric ring of AAA proteins binds a folded substrate protein that has been marked for unfolding (and eventual destruction) by a recognition tag such as a polyubiquitin chain (see below) or the peptide added to mark incompletely synthesized proteins (see Figure 6–81). A conformational change, made irreversible by ATP hydrolysis, pulls the substrate into the central core and strains the ring structure. At this point, the substrate protein, which is being tugged upon, can partially unfold and enter further into the pore or it can maintain its structure and dissociate. Very stable protein substrates may require hundreds of cycles of ATP hydrolysis and dissociation before they are successfully pulled into the AAA ring. Once unfolded, the substrate protein moves relatively quickly through the pore by successive rounds of ATP hydrolysis. (A, from X. Zhang et al., *Mol. Cell* 6:1473–1484, 2000, and A.N. Lupas and J. Martin, *Curr. Opin. Struct. Biol.* 12:746–753, 2002; B, from R.T. Sauer et al., *Cell* 119:9–18, 2004. All with permission from Elsevier.)

many different intracellular proteins. For many proteins, tagging by ubiquitin results in their destruction by the proteasome. However, in other cases, ubiquitin tagging has an entirely different meaning. Ultimately, it is the number of ubiquitin molecules added and the way in which they are linked together that determines how the cell interprets the ubiquitin message (**Figure 6–93**). In the following sections, we emphasize the role of ubiquitylation in signifying protein degradation.

An Elaborate Ubiquitin-Conjugating System Marks Proteins for Destruction

Ubiquitin is prepared for conjugation to other proteins by the ATP-dependent *ubiquitin-activating enzyme* (E1), which creates an activated, E1-bound ubiquitin that is subsequently transferred to one of a set of *ubiquitin-conjugating* (E2) enzymes (Figure 6–92B). The E2 enzymes act in conjunction with accessory (E3) proteins. In the E2–E3 complex, called *ubiquitin ligase*, the E3 component binds to specific degradation signals, called degrons, in protein substrates, helping E2 to form a *polyubiquitin* chain linked to a lysine of the substrate protein. In this chain, the C-terminal residue of each ubiquitin is linked to a specific lysine of the preceding ubiquitin molecule (see Figure 6–93), producing a linear series of ubiquitin–ubiquitin conjugates (Figure 6–92C). It is this polyubiquitin chain on a target protein that is recognized by a specific receptor in the proteasome.

There are roughly 30 structurally similar but distinct E2 enzymes in mammals, and hundreds of different E3 proteins that form complexes with specific E2 enzymes. The ubiquitin–proteasome system thus consists of many distinct but similarly organized proteolytic pathways, which have in common both the

E1 enzyme at the "top" and the proteasome at the "bottom," and differ by the compositions of their E2–E3 ubiquitin ligases and accessory factors. Distinct ubiquitin ligases recognize different degradation signals, and therefore target distinct subsets of intracellular proteins for destruction.

Denatured or otherwise misfolded proteins, as well as proteins containing oxidized or other abnormal amino acids, are recognized and destroyed because abnormal proteins tend to present on their surface amino acid sequences or conformational motifs that are recognized as degradation signals by a set of E3 molecules in the ubiquitin–proteasome system; these sequences must of course be buried and therefore inaccessible in the normal counterparts of these proteins. However, a proteolytic pathway that recognizes and destroys abnormal proteins must be able to distinguish between *completed* proteins that have "wrong" conformations and the many growing polypeptides on ribosomes (as well as polypeptides just released from ribosomes) that have not yet achieved their normal folded conformation. This is not a trivial problem; the ubiquitin–proteasome system is thought to destroy many of the nascent and newly formed protein molecules not because these proteins are abnormal as such, but because they transiently expose degradation signals that are buried in their mature (folded) state.

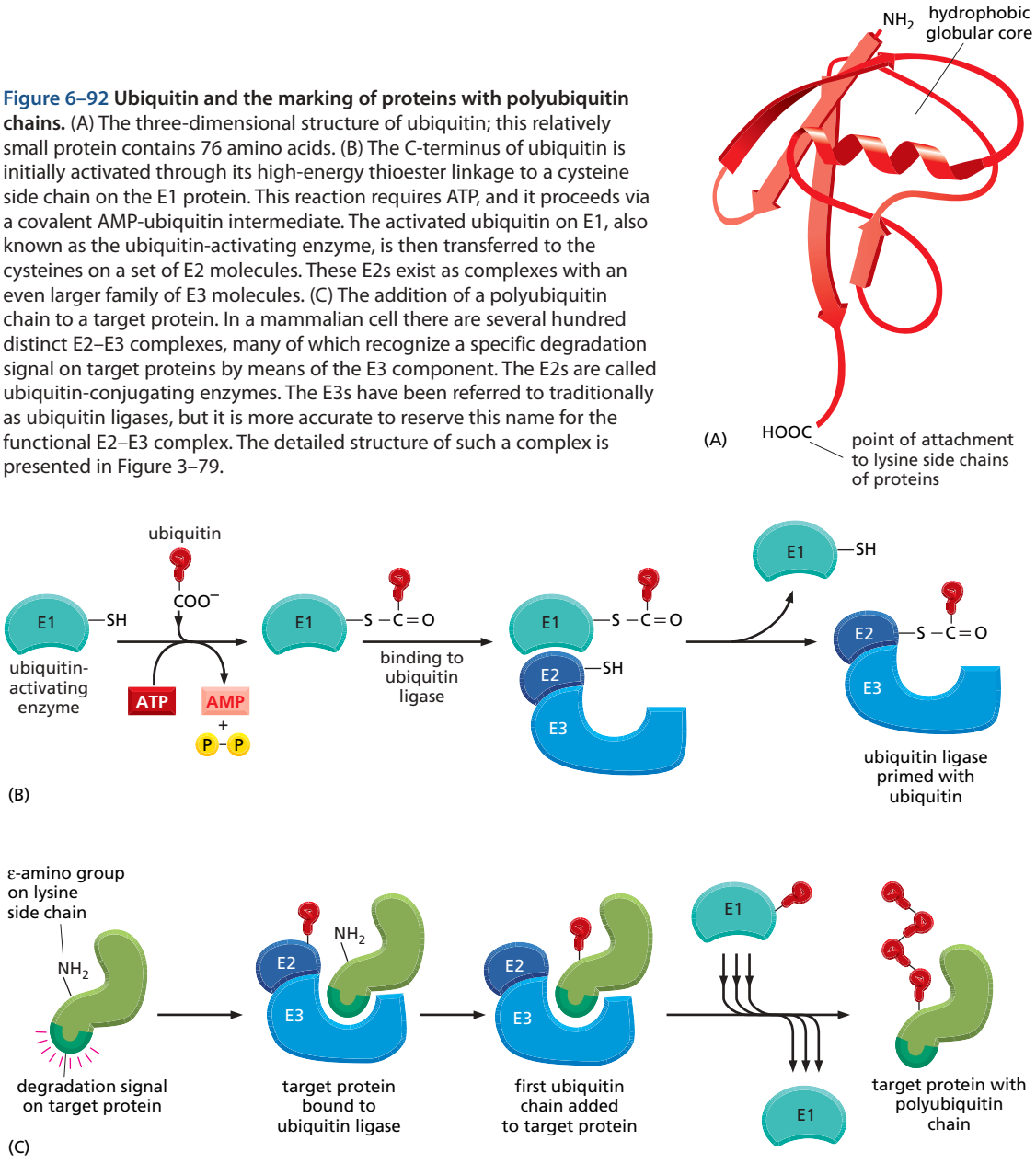


Figure 6–92 Ubiquitin and the marking of proteins with polyubiquitin chains. (A) The three-dimensional structure of ubiquitin; this relatively small protein contains 76 amino acids. (B) The C-terminus of ubiquitin is initially activated through its high-energy thioester linkage to a cysteine side chain on the E1 protein. This reaction requires ATP, and it proceeds via a covalent AMP-ubiquitin intermediate. The activated ubiquitin on E1, also known as the ubiquitin-activating enzyme, is then transferred to the cysteines on a set of E2 molecules. These E2s exist as complexes with an even larger family of E3 molecules. (C) The addition of a polyubiquitin chain to a target protein. In a mammalian cell there are several hundred distinct E2–E3 complexes, many of which recognize a specific degradation signal on target proteins by means of the E3 component. The E2s are called ubiquitin-conjugating enzymes. The E3s have been referred to traditionally as ubiquitin ligases, but it is more accurate to reserve this name for the functional E2–E3 complex. The detailed structure of such a complex is presented in Figure 3–79.

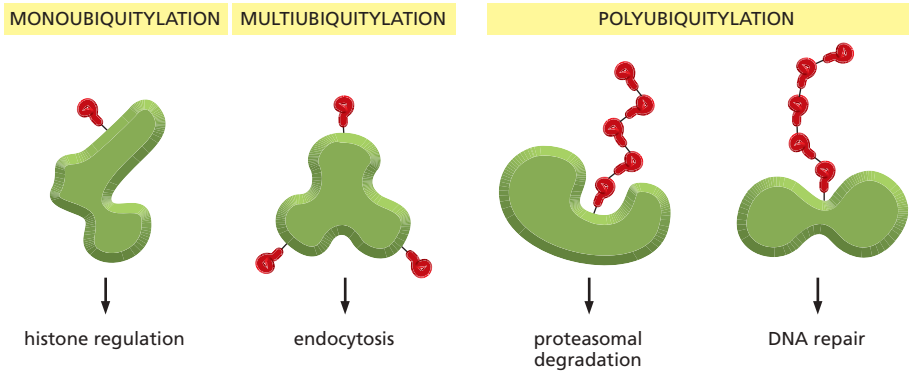


Figure 6–93 The marking of proteins by ubiquitin. Each modification pattern shown can have a specific meaning to the cell. The two types of polyubiquitylation differ in the way the ubiquitin molecules are linked together. Linkage through Lys48 signifies degradation by the proteasome whereas that through Lys63 has other meanings. Ubiquitin markings are "read" by proteins that specifically recognize each type of modification.

Many Proteins Are Controlled by Regulated Destruction

One function of intracellular proteolytic mechanisms is to recognize and eliminate misfolded or otherwise abnormal proteins, as just described. Yet another function of these proteolytic pathways is to confer short lifetimes on specific normal proteins whose concentrations must change promptly with alterations in the state of a cell. Some of these short-lived proteins are degraded rapidly at all times, while many others are *conditionally* short-lived, that is, they are metabolically stable under some conditions but become unstable upon a change in the cell's state. For example, mitotic cyclins are long-lived throughout the cell cycle until their sudden degradation at the end of mitosis, as explained in Chapter 17.

How is such a regulated destruction of a protein controlled? Several mechanisms are illustrated through specific examples that appear later in this book. In one general class of mechanism (**Figure 6–94**A), the activity of a ubiquitin ligase is turned on either by E3 phosphorylation or by an allosteric transition in an E3 protein caused by its binding to a specific small or large molecule. For example, the anaphase-promoting complex (APC) is a multisubunit ubiquitin ligase that is activated by a cell-cycle-timed subunit addition at mitosis. The activated APC then causes the degradation of mitotic cyclins and several other regulators of the metaphase–anaphase transition (see Figure 17–44).

Alternatively, in response either to intracellular signals or to signals from the environment, a degradation signal can be created in a protein, causing its rapid ubiquitylation and destruction by the proteasome. One common way to create such a signal is to phosphorylate a specific site on a protein that unmasks a normally hidden degradation signal. Another way to unmask such a signal is by the regulated dissociation of a protein subunit. Finally, powerful degradation signals can be created by cleaving a single peptide bond, provided that this cleavage creates a new N-terminus that is recognized by a specific E3 as a "destabilizing" N-terminal residue (Figure 6–94B).

The N-terminal type of degradation signal arises because of the "N-end rule," which relates the lifetime of a protein *in vivo* to the identity of its N-terminal residue. There are 12 destabilizing residues in the N-end rule of the yeast *S. cerevisiae* (Arg, Lys, His, Phe, Leu, Tyr, Trp, Ile, Asp, Glu, Asn, and Gln), out of the 20 standard amino acids. The destabilizing N-terminal residues are recognized by a special ubiquitin ligase that is conserved from yeast to humans.

As we have seen, all proteins are initially synthesized bearing methionine (or formylmethionine in bacteria), as their N-terminal residue, which is a stabilizing residue in the N-end rule. Special proteases, called methionine aminopeptidases, will often remove the first methionine of a nascent protein, but they will do so only if the second residue is also stabilizing according to N-end rule. Therefore, it was initially unclear how N-end rule substrates form *in vivo*. However, it is now understood that these substrates are formed by site-specific proteases. For example, a subunit of cohesin, a protein complex that holds sister chromatids together, is cleaved by a highly specific protease during the metaphase–anaphase transition. This cell-cycle-regulated cleavage allows separation of the sister chromatids and leads to the completion of mitosis (see

Figure 17–44). The C-terminal fragment of the cleaved subunit bears an N-terminal arginine, a destabilizing residue in the N-end rule. Mutant cells lacking the N-end rule pathway exhibit a greatly increased frequency of chromosome loss, presumably because a failure to degrade this fragment of the cohesin subunit interferes with the formation of new chromatid-associated cohesin complexes in the next cell cycle.

Abnormally Folded Proteins Can Aggregate to Cause Destructive Human Diseases

Many inherited human diseases (for example, sickle-cell anemia (see p. 1495) and α-1-antitrypsin deficiency, a condition that often leads to liver disease and emphysema) result from mutant proteins that escape the cell's quality controls, fold abnormally, and form aggregates. By absorbing critical macromolecules, these aggregates can severely damage cells and even cause cell death. Often, the inheritance of a single mutant allele of a gene can cause disease, since the normal copy of the gene cannot protect the cell from the destructive properties of the aggregate.

In normal humans, the gradual decline of the cell's protein quality controls can also cause disease by permitting normal proteins to form aggregates (**Figure 6–95**). In some cases, the protein aggregates are released from dead cells and accumulate in the extracellular matrix that surrounds the cells in a tissue, and in extreme cases they can also damage tissues. Because the brain is composed of a highly organized collection of nerve cells, it is especially vulnerable.

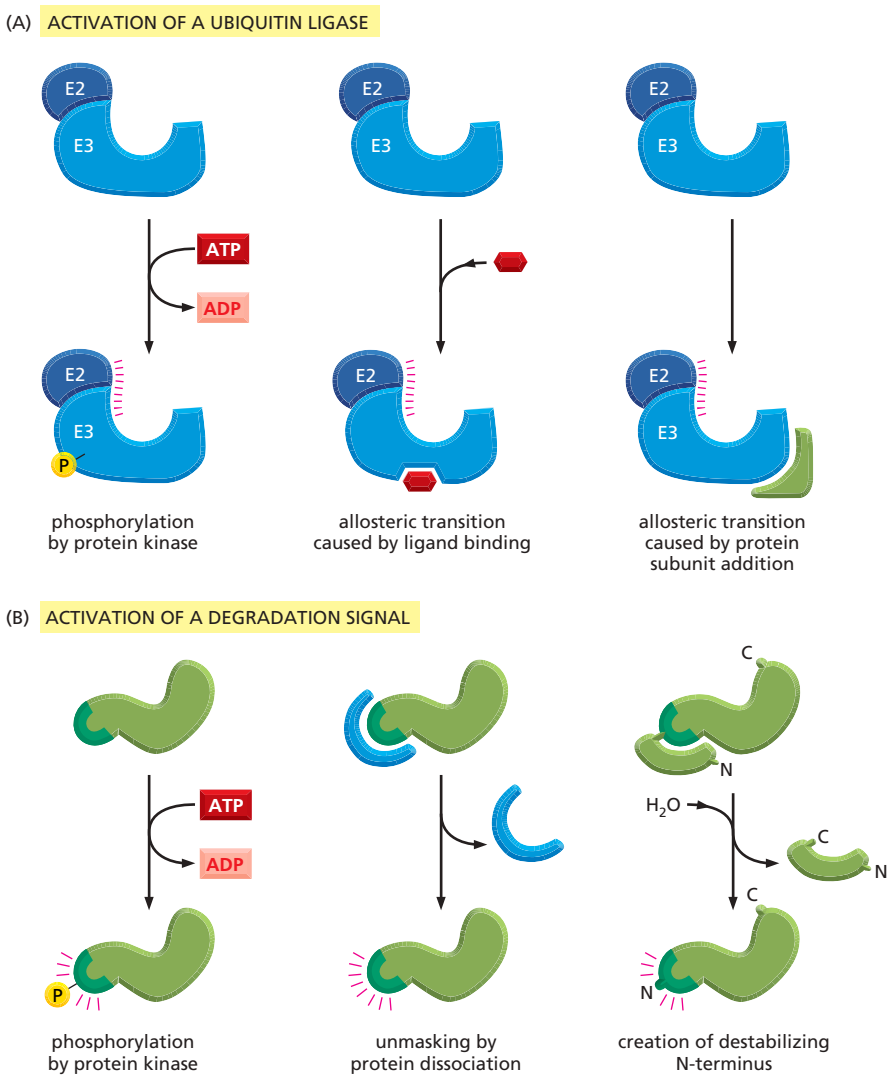


Figure 6–94 Two general ways of inducing the degradation of a specific protein. (A) Activation of a specific E3 molecule creates a new ubiquitin ligase. (B) Creation of an exposed degradation signal in the protein to be degraded. This signal binds a ubiquitin ligase, causing the addition of a polyubiquitin chain to a nearby lysine on the target protein. All six pathways shown are known to be used by cells to induce the movement of selected proteins into the proteasome.

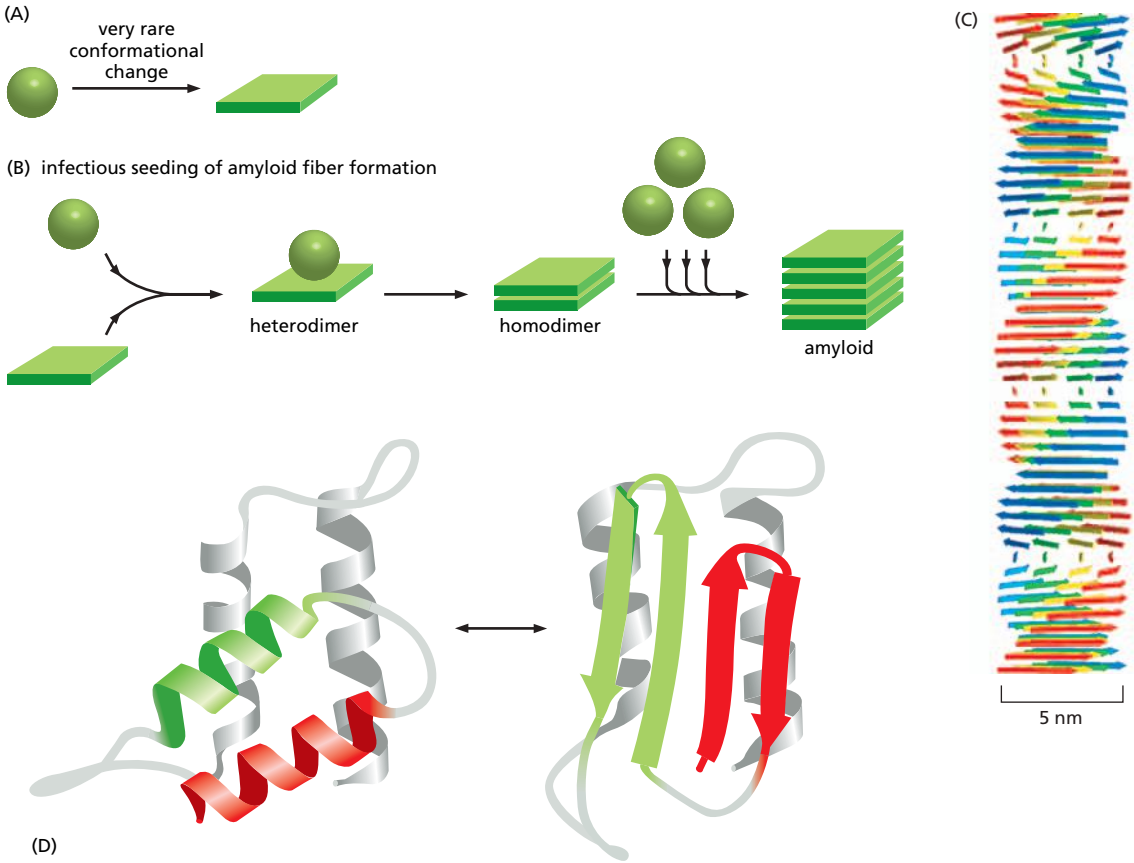


Figure 6–95 Protein aggregates that cause human disease. (A) Schematic illustration of the type of conformational change in a protein that produces material for a cross-beta filament. (B) Diagram illustrating the self-infectious nature of the protein aggregation that is central to prion diseases. PrP (prion protein) is highly unusual because the misfolded version of the protein, called PrP*, induces the normal PrP protein it contacts to change its conformation, as shown. Most of the human diseases caused by protein aggregation are caused by the overproduction of a variant protein that is especially prone to aggregation, but the protein aggregate cannot spread from one animal to another. (C) Drawing of a cross-beta filament, a common type of protease-resistant protein aggregate found in many human neurological diseases. Because the hydrogen-bond interactions in a β sheet form between polypeptide backbone atoms (see Figure 3–9), a number of different abnormally folded proteins can produce this structure. (D) One of several possible models for the conversion of PrP to PrP*, showing the likely change of two α helices into four β strands. Although the structure of the normal protein has been determined accurately, the structure of the infectious form is not yet known with certainty because the aggregation has prevented the use of standard structural techniques. (C, courtesy of Louise Serpell, adapted from M. Sunde et al., *J. Mol. Biol.* 273:729–739, 1997. With permission from Academic Press; D, adapted from S.B. Prusiner, *Trends Biochem. Sci.* 21:482–487, 1996. With permission from Elsevier.)

Not surprisingly, therefore, protein aggregates primarily cause neurodegenerative diseases. Prominent among these are Huntington's disease and Alzheimer's disease—the latter causing age-related dementia in more than 20 million people in today's world.

For a particular type of protein aggregate to survive, grow, and damage an organism, it must be highly resistant to proteolysis both inside and outside the cell. Many of the protein aggregates that cause problems form fibrils built from a series of polypeptide chains that are layered one over the other as a continuous stack of β sheets. This so-called *cross-beta filament* (Figure 6–95C), a structure particularly resistant to proteolysis, is observed in many of the neurological disorders caused by protein aggregates, where it produces distinctly staining deposits known as *amyloids*.

One particular variety of these pathologies has attained special notoriety. These are the **prion diseases**. Unlike Huntington's or Alzheimer's, prion diseases can spread from one organism to another, providing that the second organism eats a tissue containing the protein aggregate. A set of diseases—called scrapie in sheep, Creutzfeldt–Jacob disease (CJD) in humans, and bovine spongiform encephalopathy (BSE) in cattle—are caused by a misfolded, aggregated form of a protein called PrP (for prion protein). The PrP is normally located on the outer surface of the plasma membrane, most prominently in neurons. Its normal

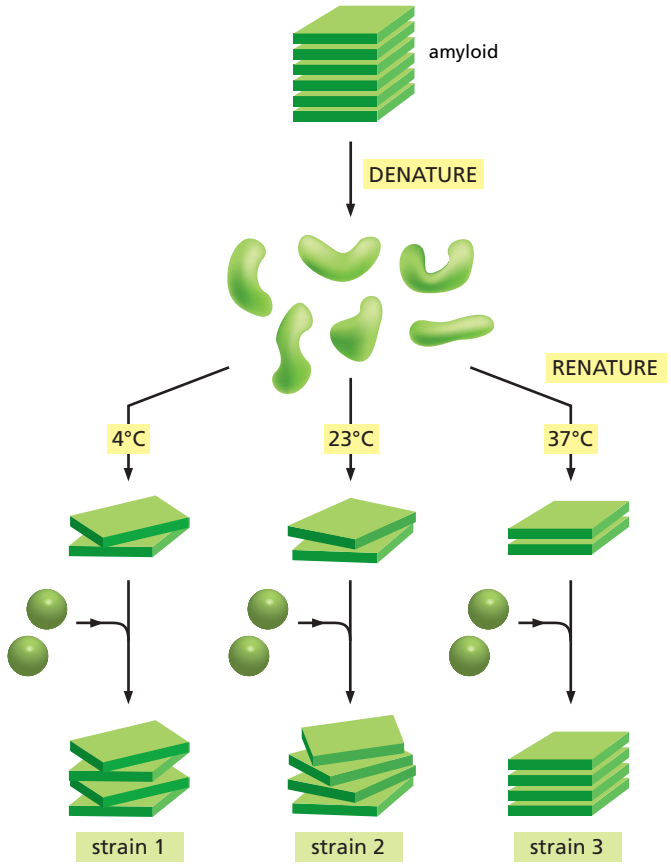


Figure 6–96 Creation of different prion strains *in vitro*. In this experiment, amyloid fibers were denatured and the components renatured at different temperatures. This treatment produced three distinctive types of amyloids, each of which could self-propagate when new subunits are added.

function is not known. However, PrP has the unfortunate property of being convertible to a very special abnormal conformation (see Figure 6–95A). This conformation not only forms protease-resistant, cross-beta filaments; it is also "infectious" because it converts normally folded molecules of PrP to the same pathological form. This property creates a positive feedback loop that propagates the abnormal form of PrP, called PrP* (see Figure 6–95B) and thereby allows the pathological conformation to spread rapidly from cell to cell in the brain, eventually causing death in both animals and humans. It can be dangerous to eat the tissues of animals that contain PrP*, as witnessed by the spread of BSE (commonly referred to as "mad cow disease") from cattle to humans in Great Britain. Fortunately, in the absence of PrP*, PrP is extraordinarily difficult to convert to its abnormal form.

Although very few proteins have the potential to misfold into an infectious conformation, another example causes an otherwise mysterious "protein-only inheritance" observed in yeast cells. The ability to study infectious proteins in yeast has clarified another remarkable feature of prions. These protein molecules can form several distinctively different types of aggregates from the same polypeptide chain. Moreover, each type of aggregate can be infectious, forcing normal protein molecules to adopt the same type of abnormal structure. Thus, several different "strains" of infectious particles can arise from the same polypeptide chain (**Figure 6–96**). How a single polypeptide sequence can adopt multiple aggregate forms is not fully understood; it is possible that all prion aggregates resemble cross-beta filaments (see Figure 6–95C) where the structure is held together predominantly with main peptide chain interactions. This would leave the amino acid side chains free to adopt different conformations and, if the structures are self-propagating, the existence of different strains could be explained.

Finally, although prions were discovered because they cause disease, they also appear to have some positive roles in the cell. For example, some species of fungi use prion transformations to establish different types of cells. Although the idea is controversial, it has even been proposed that prions have a role in consolidating memories in complex, multicellular organisms like ourselves.

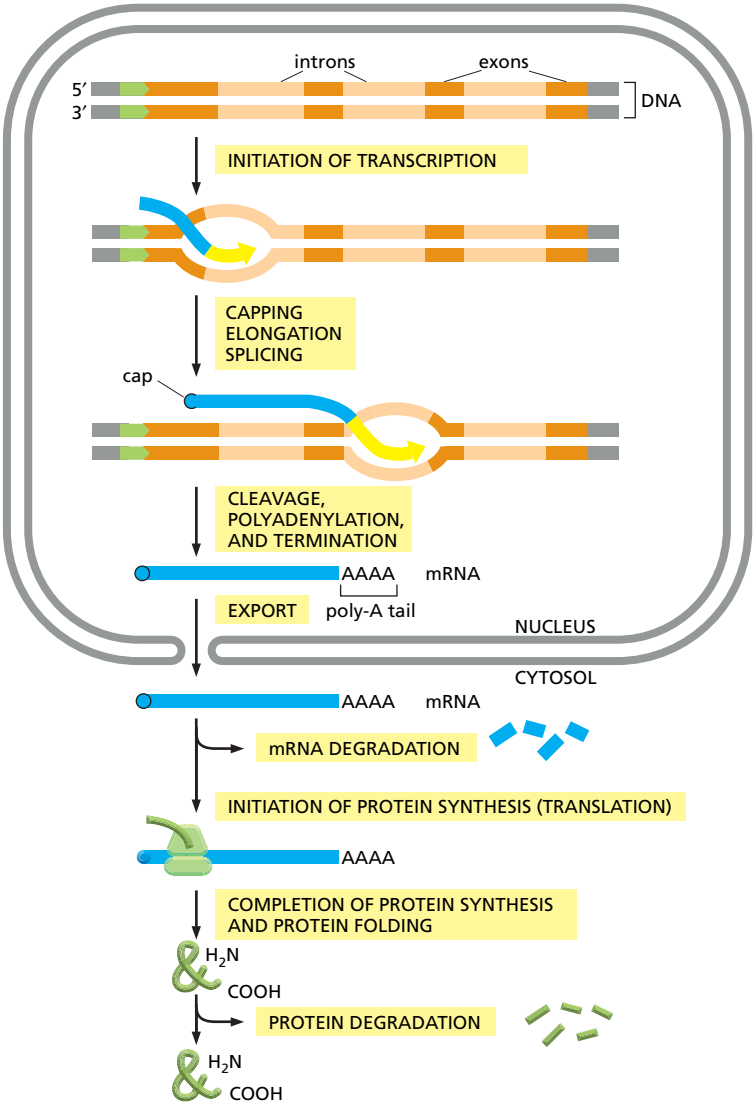


Figure 6–97 The production of a protein by a eucaryotic cell. The final level of each protein in a eucaryotic cell depends upon the efficiency of each step depicted.

There Are Many Steps From DNA to Protein

We have seen so far in this chapter that many different types of chemical reactions are required to produce a properly folded protein from the information contained in a gene (**Figure 6–97**). The final level of a properly folded protein in a cell therefore depends upon the efficiency with which each of the many steps is performed.

In the following chapter, we shall see that cells have the ability to change the levels of their proteins according to their needs. In principle, any or all of the steps in Figure 6–97 could be regulated for each individual protein. As we shall see in Chapter 7, there are examples of regulation at each step from gene to protein. However, the initiation of transcription is the most common point for a cell to regulate the expression of each of its genes. This makes sense, inasmuch as the most efficient way to keep a gene from being expressed is to block the very first step—the transcription of its DNA sequence into an RNA molecule.

Summary

The translation of the nucleotide sequence of an mRNA molecule into protein takes place in the cytoplasm on a large ribonucleoprotein assembly called a ribosome. The amino acids used for protein synthesis are first attached to a family of tRNA molecules, each of which recognizes, by complementary base-pair interactions, particular sets of three nucleotides in the mRNA (codons). The sequence of nucleotides in the mRNA is then read from one end to the other in sets of three according to the genetic code.

To initiate translation, a small ribosomal subunit binds to the mRNA molecule at a start codon (AUG) that is recognized by a unique initiator tRNA molecule. A large ribosomal subunit binds to complete the ribosome and begin protein synthesis. During this phase, aminoacyl-tRNAs—each bearing a specific amino acid—bind sequentially to the appropriate codons in mRNA through complementary base pairing between tRNA anticodons and mRNA codons. Each amino acid is added to the C-terminal end of the growing polypeptide in four sequential steps: aminoacyl-tRNA binding, followed by peptide bond formation, followed by two ribosome translocation steps. Elongation factors use GTP hydrolysis to drive these reactions forward and to improve the accuracy of amino acid selection. The mRNA molecule progresses codon by codon through the ribosome in the 5′-to-3′ direction until it reaches one of three stop codons. A release factor then binds to the ribosome, terminating translation and releasing the completed polypeptide.

Eucaryotic and bacterial ribosomes are closely related, despite differences in the number and size of their rRNA and protein components. The rRNA has the dominant role in translation, determining the overall structure of the ribosome, forming the binding sites for the tRNAs, matching the tRNAs to codons in the mRNA, and creating the active site of the peptidyl transferase enzyme that links amino acids together during translation.

In the final steps of protein synthesis, two distinct types of molecular chaperones guide the folding of polypeptide chains. These chaperones, known as Hsp60 and Hsp70, recognize exposed hydrophobic patches on proteins and serve to prevent the protein aggregation that would otherwise compete with the folding of newly synthesized proteins into their correct three-dimensional conformations. This protein folding process must also compete with an elaborate quality control mechanism that destroys proteins with abnormally exposed hydrophobic patches. In this case, ubiquitin is covalently added to a misfolded protein by a ubiquitin ligase, and the resulting polyubiquitin chain is recognized by the cap on a proteasome that moves the entire protein to the interior of the proteasome for proteolytic degradation. A closely related proteolytic mechanism, based on special degradation signals recognized by ubiquitin ligases, is used to determine the lifetimes of many normally folded proteins. By this method, selected normal proteins are removed from the cell in response to specific signals.

THE RNA WORLD AND THE ORIGINS OF LIFE

We have seen that the expression of hereditary information requires extraordinarily complex machinery and proceeds from DNA to protein through an RNA intermediate. This machinery presents a central paradox: if nucleic acids are required to synthesize proteins and proteins are required, in turn, to synthesize nucleic acids, how did such a system of interdependent components ever arise? One view is that an *RNA world* existed on Earth before modern cells arose (**Figure 6–98**). According to this hypothesis, RNA both stored genetic information and catalyzed the chemical reactions in primitive cells. Only later in evolutionary time did DNA take over as the genetic material and proteins become the major catalyst and structural component of cells. If this idea is correct, then the transition out of the RNA world was never complete; as we have seen in this chapter, RNA still catalyzes several fundamental reactions in modern-day cells, which can be viewed as molecular fossils of an earlier world.

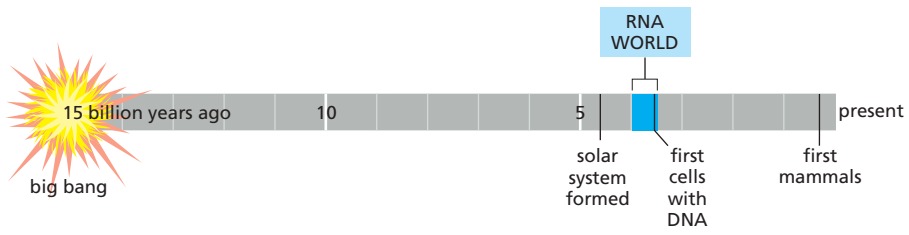


Figure 6–98 Time line for the universe, suggesting the early existence of an RNA world of living systems.

In this section we present some of the arguments in support of the RNA world hypothesis. We will see that several of the more surprising features of modern-day cells, such as the ribosome and the pre-mRNA splicing machinery, are most easily explained by viewing them as descendants of a complex network of RNA-mediated interactions that dominated cell metabolism in the RNA world. We also discuss how DNA may have taken over as the genetic material, how the genetic code may have arisen, and how proteins may have eclipsed RNA to perform the bulk of biochemical catalysis in modern-day cells.

Life Requires Stored Information

It has been proposed that the first "biological" molecules on Earth were formed by metal-based catalysis on the crystalline surfaces of minerals. In principle, an elaborate system of molecular synthesis and breakdown (metabolism) could have existed on these surfaces long before the first cells arose. Although controversial, many scientists believe that an extensive phase of "chemical evolution" took place on the prebiotic Earth, during which small molecules that could catalyze their own synthesis competed with each other for raw materials.

But life requires much more than this. As described in Chapter 1, *heredity* is perhaps the central feature of life. Not only must a cell use raw materials to create a network of catalyzed reactions, it must do so according to an elaborate set of instructions encoded in the hereditary information. The replication of this information ensures that the complex metabolism of cells can accurately reproduce itself. Another crucial feature of life is the genetic variability that results from changes in the hereditary information. This variability, acted upon by selective pressures, is responsible for the great diversity of life on our planet.

Thus, the emergence of life requires a way to store information, a way to duplicate it, a way to change it, and a way to convert the information through catalysis into favorable chemical reactions. But how could such a system begin to be formed? In present-day cells the most versatile catalysts are polypeptides, composed of many different amino acids with chemically diverse side chains and, consequently, able to adopt diverse three-dimensional forms that bristle with reactive chemical groups. Polypeptides also carry information, in the order of their amino acid subunits. But there is no known way in which a polypeptide can reproduce itself by directly specifying the formation of another of precisely the same sequence.

Polynucleotides Can Both Store Information and Catalyze Chemical Reactions

Polynucleotides have one property that contrasts with those of polypeptides: they can directly guide the formation of copies of their own sequence. This capacity depends on complementary base pairing of nucleotide subunits, which enables one polynucleotide to act as a template for the formation of another. As we have seen in this and the preceding chapter, such complementary templating mechanisms lie at the heart of DNA replication and transcription in modern-day cells.

But the efficient synthesis of polynucleotides by such complementary templating mechanisms requires catalysts to promote the polymerization reaction: without catalysts, polymer formation is slow, error-prone, and inefficient. Today, template-based nucleotide polymerization is rapidly catalyzed by protein enzymes—such as the DNA and RNA polymerases. How could such polymerization be catalyzed before proteins with the appropriate enzymatic specificity existed? The beginnings of an answer to this question came from the discovery in 1982 that RNA molecules themselves can act as catalysts. We have seen in this chapter, for example, that a molecule of RNA catalyzes one of the central reactions in the cell, the covalent joining of amino acids to form proteins. The unique potential of RNA molecules to act both as information carrier and as catalyst forms the basis of the RNA world hypothesis.

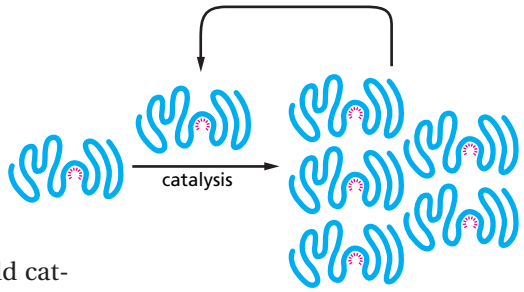


Figure 6–99 An RNA molecule that can catalyze its own synthesis. This hypothetical process would require catalysis of the production of both a second RNA strand of complementary nucleotide sequence and the use of this second RNA molecule as a template to form many molecules of RNA with the original sequence. The *red* rays represent the active site of this hypothetical RNA enzyme.

RNA therefore has all the properties required of a molecule that could catalyze a variety of chemical reactions, including those that lead to its own synthesis (**Figure 6–99**). Although self-replicating systems of RNA molecules have not been found in nature, scientists are confident that they can be constructed in the laboratory. While this demonstration would not prove that self-replicating RNA molecules were essential in the origin of life on Earth, it would certainly indicate that such a scenario is possible.

A Pre-RNA World May Predate the RNA World

Although RNA seems well suited to form the basis for a self-replicating set of biochemical catalysts, it is not clear that RNA was the first kind of molecule to do so. From a purely chemical standpoint, it is difficult to imagine how long RNA molecules could be formed initially by purely nonenzymatic means. For one thing, the precursors of RNA, the ribonucleotides, are difficult to form nonenzymatically. Moreover, the formation of RNA requires that a long series of 3′-to-5′ phosphodiester linkages assemble in the face of a set of competing reactions, including hydrolysis, 2′-to-5′ linkages, and 5′-to-5′ linkages. Given these problems, it has been suggested that the first molecules to possess both catalytic activity and information storage capabilities may have been polymers that resemble RNA but are chemically simpler (**Figure 6–100**). We do not have any remnants of these compounds in present-day cells, nor do such compounds leave fossil records. Nonetheless, the relative simplicity of these "RNA-like polymers" suggests that one of them, rather than RNA itself, may have been the first biopolymer on Earth capable of both information storage and catalytic activity.

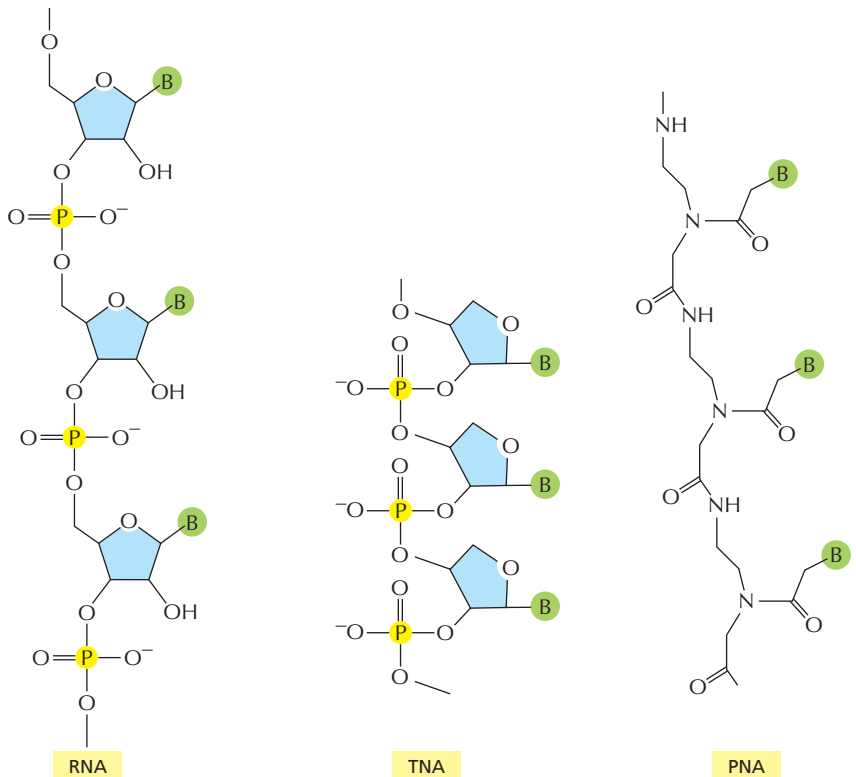


Figure 6–100 Structures of RNA and two related information-carrying polymers. In each case, B indicates a purine or pyrimidine base. The polymer TNA (threose nucleic acid) has a 4-carbon sugar unit in contrast to the 5-carbon ribose in RNA. In PNA (peptide nucleic acid), the ribose phosphate backbone of RNA has been replaced by the peptide backbone found in proteins. Like RNA, TNA and PNA can form double helices through complementary base-pairing, and each could therefore in principle serve as a template for its own synthesis.

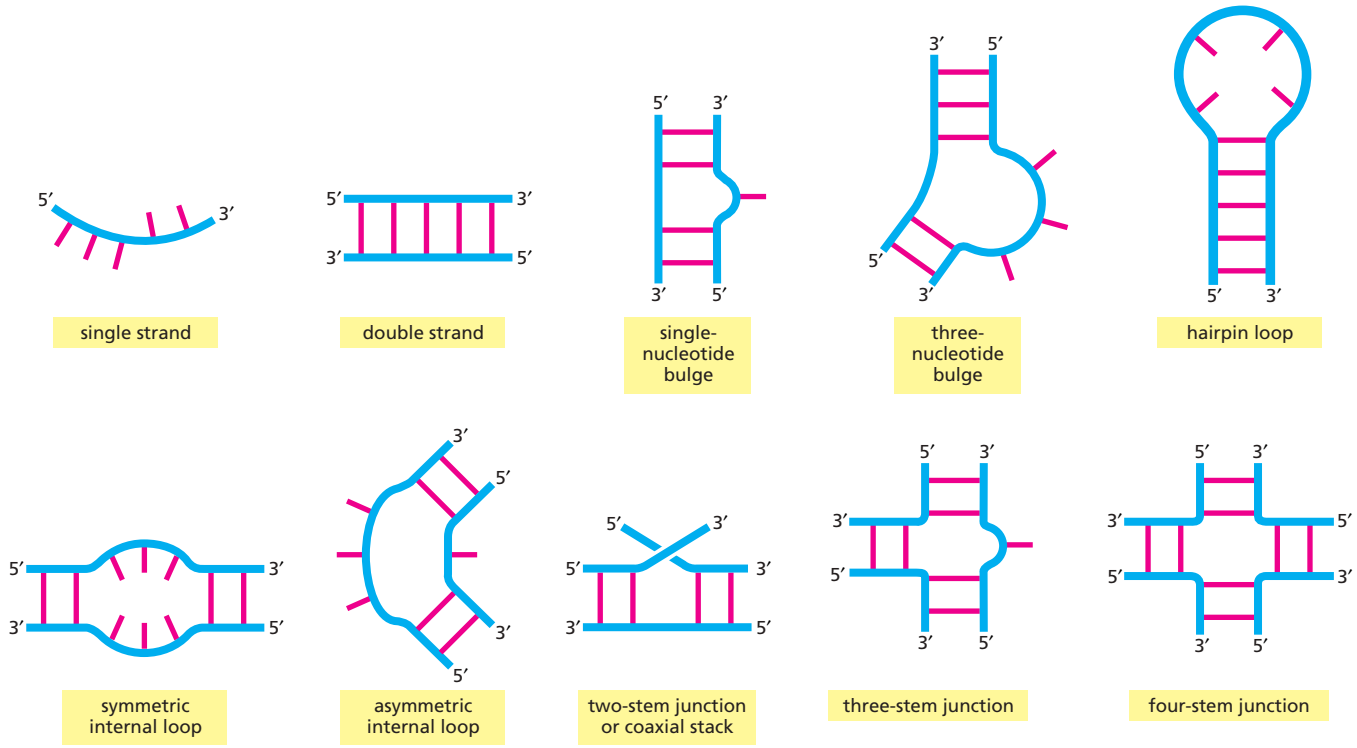


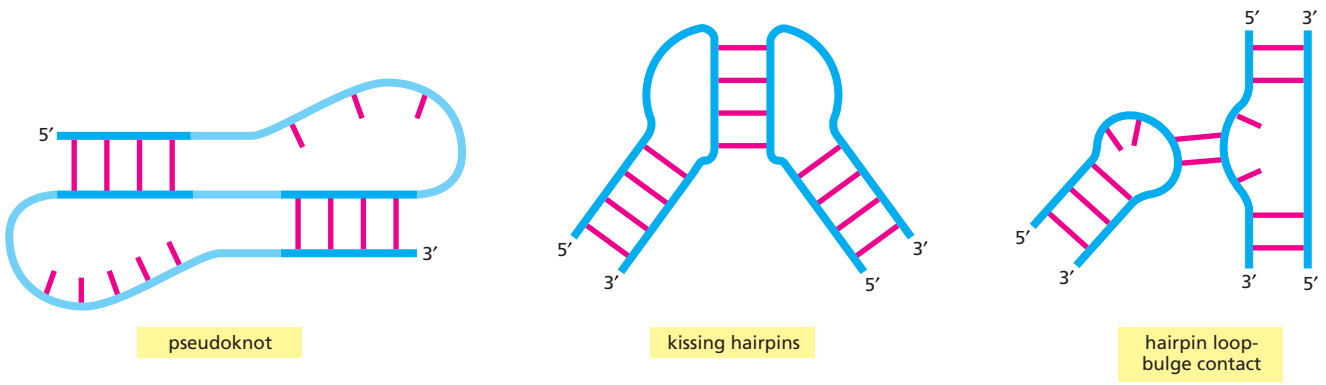
Figure 6–101 Common elements of RNA secondary structure. Conventional, complementary base-pairing interactions are indicated by *red* "rungs" in double-helical portions of the RNA.

If the pre-RNA world hypothesis is correct, then a transition to the RNA world must have occurred, presumably through the synthesis of RNA using one of these simpler polymers as both template and catalyst. While the details of the pre-RNA and RNA worlds will likely remain unknown, we know for certain that RNA molecules can catalyze a wide variety of chemical reactions, and we now turn to the properties of RNA that make this possible.

Single-Stranded RNA Molecules Can Fold into Highly Elaborate Structures

We have seen that complementary base-pairing and other types of hydrogen bonds can occur between nucleotides in the same chain, causing an RNA molecule to fold up in a unique way determined by its nucleotide sequence (see, for example, Figures 6–6, 6–52, and 6–69). Comparisons of many RNA structures have revealed conserved motifs, short structural elements that are used over and over again as parts of larger structures. **Figure 6–101** shows some of these RNA secondary structural motifs, and **Figure 6–102** shows a few common examples of more complex and often longer-range interactions, known as RNA tertiary interactions.

Figure 6–102 Examples of RNA tertiary interactions. Some of these interactions can join distant parts of the same RNA molecule or bring two separate RNA molecules together.



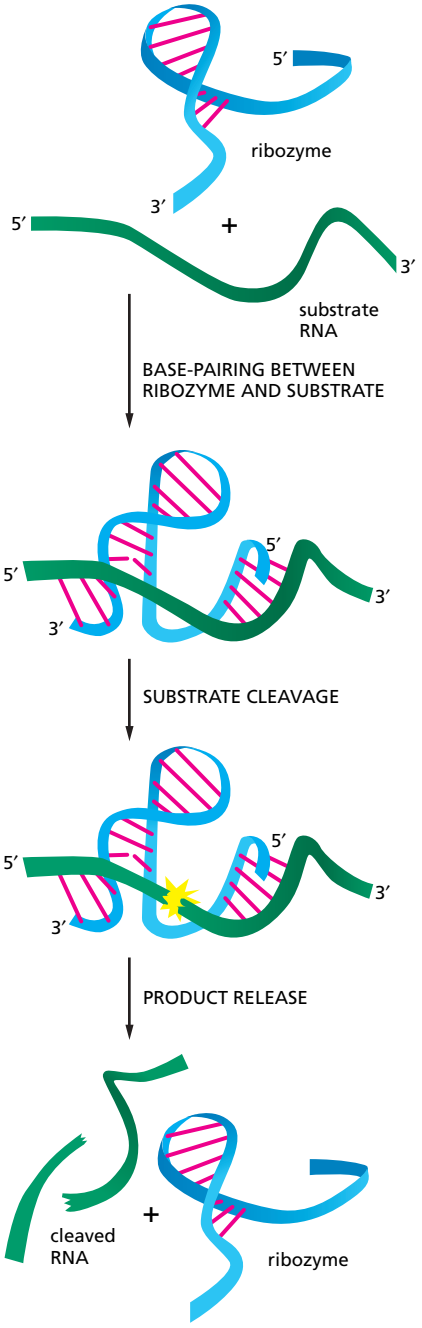


Figure 6–103 A ribozyme. This simple RNA molecule catalyzes the cleavage of a second RNA at a specific site. This ribozyme is found embedded in larger RNA genomes—called viroids—which infect plants. The cleavage, which occurs in nature at a distant location on the same RNA molecule that contains the ribozyme, is a step in the replication of the viroid genome. Although not shown in the figure, the reaction requires a Mg molecule positioned at the active site. (Adapted from T.R. Cech and O.C. Uhlenbeck, *Nature* 372:39–40, 1994. With permission from Macmillan Publishers Ltd.)

Protein catalysts require a surface with unique contours and chemical properties on which a given set of substrates can react (discussed in Chapter 3). In exactly the same way, an RNA molecule with an appropriately folded shape can serve as an enzyme (**Figure 6–103**). Like some proteins, many of these ribozymes work by positioning metal ions at their active sites. This feature gives them a wider range of catalytic activities than the limited chemical groups of the polynucleotide chain.

Relatively few catalytic RNAs are known to exist in modern-day cells, however, and much of our inference about the RNA world has come from experiments in which large pools of RNA molecules of random nucleotide sequences are generated in the laboratory. Those rare RNA molecules with a property specified by the experimenter are then selected out and studied (**Figure 6–104**). Such experiments have created RNAs that can catalyze a wide variety of biochemical reactions (**Table 6–5**), with reaction rate enhancements approaching those of proteins. Given these findings, it is not clear why protein catalysts greatly outnumber ribozymes in modern cells. Experiments have shown that RNA molecules may have more difficulty than proteins in binding to flexible, hydrophobic substrates; moreover, the availability of 20 types of amino acids over four types of bases may provide proteins with a greater number of catalytic strategies.

Like proteins, RNAs can undergo conformational changes, either in response to small molecules or to other RNAs. We saw several examples of this in the ribosome and the spliceosome, and we will see others in Chapter 7 when we discuss *riboswitches*. One of the most dramatic RNA conformational changes has been observed with an artificial ribozyme which can exist in two entirely different conformations, each with a different catalytic activity (**Figure 6–105**). Since the discovery of catalysis by RNA, it has become clear that RNA is an enormously versatile molecule, and it is therefore not unreasonable to contemplate the past existence of an RNA world with a very high level of biochemical sophistication.

Self-Replicating Molecules Undergo Natural Selection

The three-dimensional folded structure of a polynucleotide affects its stability, its actions on other molecules, and its ability to replicate. Therefore, certain polynucleotides will be especially successful in any self-replicating mixture. Because errors inevitably occur in any copying process, new variant sequences of these polynucleotides will be generated over time.

Certain catalytic activities would have had a cardinal importance in the early evolution of life. Consider in particular an RNA molecule that helps to catalyze the process of templated polymerization, taking any given RNA molecule as a template (**Figure 6–106**). Such a molecule, by acting on copies of itself, can replicate. At the same time, it can promote the replication of other types of RNA molecules in its neighborhood (**Figure 6–107**). If some of these neighboring RNAs have catalytic actions that help the survival of RNA in other ways (catalyzing ribonucleotide production, for example), a set of different types of RNA molecules, each specialized for a different activity, could evolve into a cooperative system that replicates with unusually great efficiency.

But for any of these cooperative systems to evolve, they must be present together in a compartment. For example, a set of mutually beneficial RNAs (such

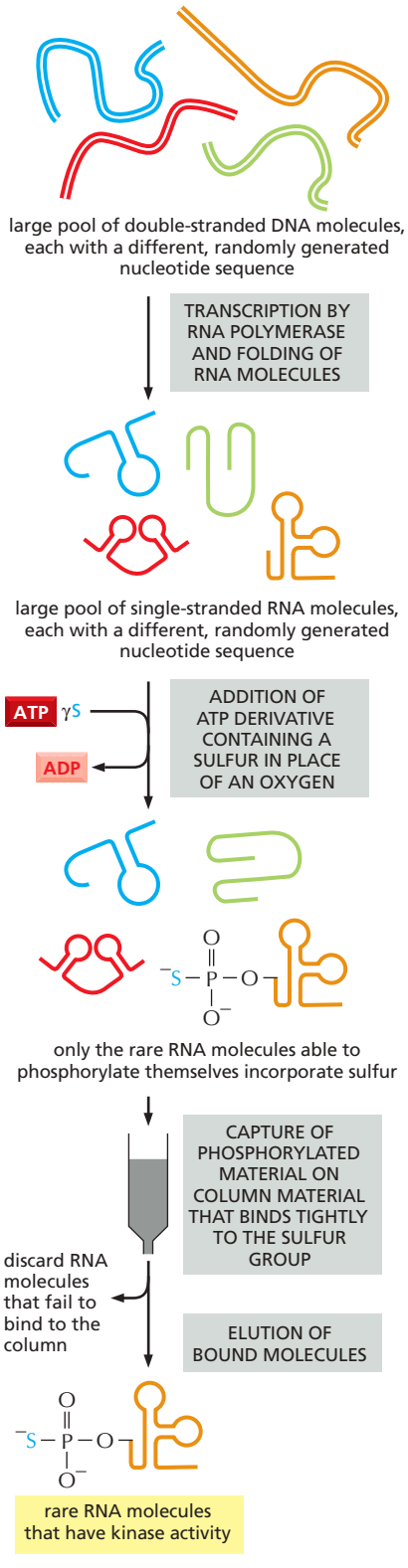


Figure 6–104 *In vitro* selection of a synthetic ribozyme. Beginning with a large pool of nucleic acid molecules synthesized in the laboratory, those rare RNA molecules that possess a specified catalytic activity can be isolated and studied. Although a specific example (that of an autophosphorylating ribozyme) is shown, variations of this procedure have been used to generate many of the ribozymes listed in Table 6–5. During the autophosphorylation step, the RNA molecules are kept sufficiently dilute to prevent the "cross"-phosphorylation of additional RNA molecules. In reality, several repetitions of this procedure are necessary to select the very rare RNA molecules with this catalytic activity. Thus, the material initially eluted from the column is converted back into DNA, amplified many fold (using reverse transcriptase and PCR as explained in Chapter 8), transcribed back into RNA, and subjected to repeated rounds of selection. (Adapted from J.R. Lorsch and J.W. Szostak, *Nature* 371:31–36, 1994. With permission from Macmillan Publishers Ltd.)

as those of Figure 6–107) could replicate themselves only if all the RNAs remained in the neighborhood of the RNA that is specialized for templated polymerization. Moreover, compartmentalization would bar parasitic RNA molecules from entering the system. Selection of a set of RNA molecules according to the quality of the self-replicating systems they generated could not therefore occur efficiently until some form of compartment evolved to contain them.

An early, crude form of compartmentalization may have been simple adsorption on surfaces or particles. The need for more sophisticated types of containment is easily fulfilled by a class of small molecules that has the simple physicochemical property of being *amphiphilic*, that is, consisting of one part that is hydrophobic (water insoluble) and another part that is hydrophilic (water soluble). When such molecules are placed in water, they aggregate, arranging their hydrophobic portions as much in contact with one another as possible and their hydrophilic portions in contact with the water. Amphiphilic molecules of appropriate shape aggregate spontaneously to form *bilayers*, creating small closed vesicles whose aqueous contents are isolated from the external medium (**Figure 6–108**). The phenomenon can be demonstrated in a test tube by simply mixing phospholipids and water together: under appropriate conditions, small vesicles will form. All present-day cells are surrounded by a *plasma membrane* consisting of amphiphilic molecules—mainly phospholipids—in this configuration; we discuss these molecules in detail in Chapter 10.

The spontaneous assembly of a set of amphiphilic molecules, enclosing a self-replicating mixture of RNAs (or pre-RNAs) and other molecules (**Figure**

Table 6–5 Some Biochemical Reactions That Can Be Catalyzed by Ribozymes

ACTIVITY	RIBOZYMES
Peptide bond formation in protein synthesis	ribosomal RNA
RNA cleavage, RNA ligation	self-splicing RNAs; RNase P; also *in vitro* selected RNA
DNA cleavage	self-splicing RNAs
RNA splicing	self-splicing RNAs, perhaps RNAs of the spliceosome
RNA polymerizaton	*in vitro* selected RNA
RNA and DNA phosphorylation	*in vitro* selected RNA
RNA aminoacylation	*in vitro* selected RNA
RNA alkylation	*in vitro* selected RNA
Amide bond formation	*in vitro* selected RNA
Glycosidic bond formation	*in vitro* selected RNA
Oxidation/reduction reactions	*in vitro* selected RNA
Carbon–carbon bond formation	*in vitro* selected RNA
Phosphoamide bond formation	*in vitro* selected RNA
Disulfide exchange	*in vitro* selected RNA

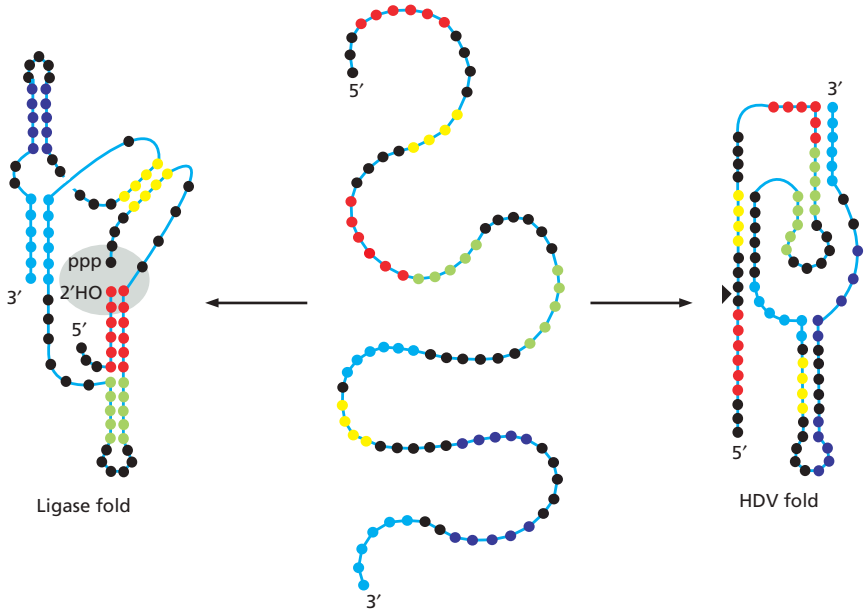


Figure 6–105 An RNA molecule that folds into two different ribozymes. This 88-nucleotide RNA, created in the laboratory, can fold into a ribozyme that carries out a self-ligation reaction *(left)* or a self-cleavage reaction *(right)*. The ligation reaction forms a 2′,5′ phosphodiester linkage with the release of pyrophosphate. This reaction seals the gap *(gray* shading), which was experimentally introduced into the RNA molecule. In the reaction carried out by the HDV fold, the RNA is cleaved at this same position, indicated by the *arrowhead*. This cleavage resembles that used in the life cycle of HDV, a hepatitis B satellite virus, hence the name of the fold. Each nucleotide is represented by a *colored dot*, with the colors used simply to clarify the two different folding patterns. The folded structures illustrate the secondary structures of the two ribozymes with regions of base-pairing indicated by close oppositions of the *colored dots*. Note that the two ribozyme folds have no secondary structure in common. (Adapted from E.A. Schultes and D.P. Bartel, *Science* 289:448–452, 2000. With permission from AAAS.)

6–109), presumably formed the first membrane-bounded cells. Although it is not clear at what point in the evolution of biological catalysts this might have occurred, once RNA molecules were sealed within a closed membrane they could begin to evolve in earnest as carriers of genetic instructions: new variants could be selected not merely on the basis of their own structure, but also according to their effect on the other molecules in the same compartment. The nucleotide sequences of the RNA molecules could now be expressed in the character of a unitary living cell.

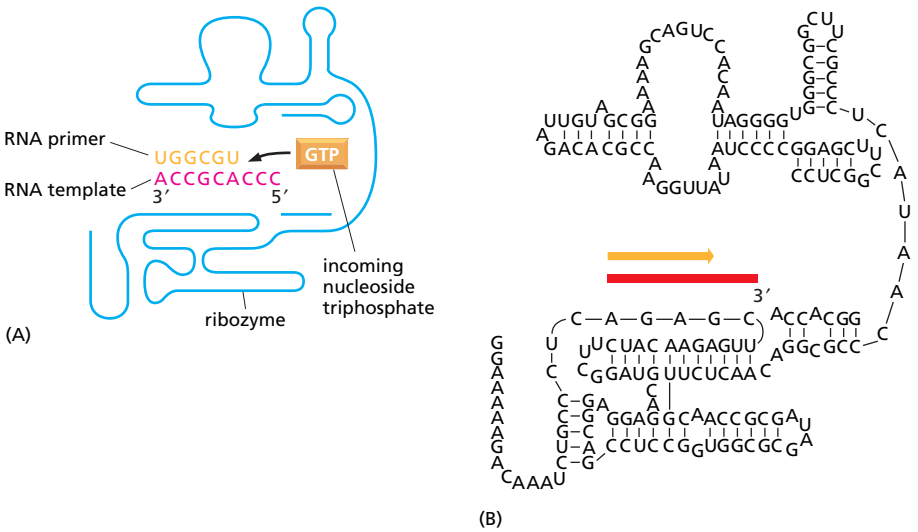


Figure 6–106 A ribozyme created in the laboratory that can catalyze templated synthesis of RNA from nucleoside triphosphates. (A) Schematic diagram of the ribozyme showing one step of the templated polymerization reaction it catalyzes. (B) Nucleotide sequence of the ribozyme with base pairings indicated. Although relatively inefficient (it can only synthesize short lengths of RNA), this ribozyme adds the correct base, as specified by the template, over 95% of the time. (From W.K. Johnston et al., *Science* 292:1319–1325, 2001. With permission from AAAS.)

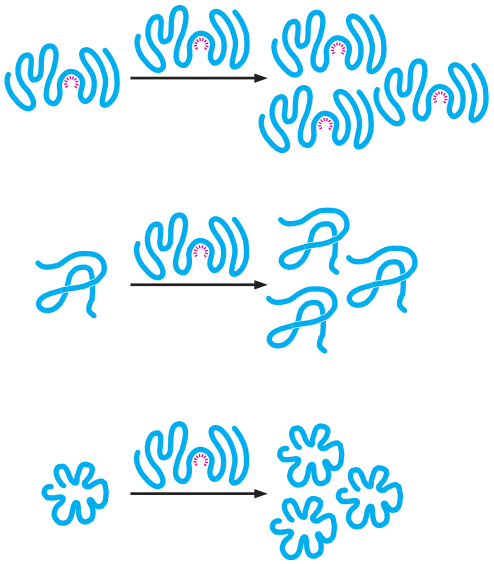


Figure 6–107 A family of mutually supportive RNA molecules. One molecule is a ribozyme that replicates itself as well as the other RNA molecules. The other molecules would catalyze secondary tasks needed for the survival of the cooperative system, for example, by synthesizing ribonucleotides for RNA synthesis or phospholipids for compartmentalization.

How Did Protein Synthesis Evolve?

The molecular processes underlying protein synthesis in present-day cells seem inextricably complex. Although we understand most of them, they do not make conceptual sense in the way that DNA transcription, DNA repair, and DNA replication do. It is especially difficult to imagine how protein synthesis evolved because it is now performed by a complex interlocking system of protein and RNA molecules; obviously the proteins could not have existed until an early version of the translation apparatus was already in place. The RNA world hypothesis is especially appealing because the use of RNA in both information and catalysis seems both economic and conceptually simple. As attractive as this idea is for envisioning early life, it does not explain how the modern-day system of protein synthesis arose. Although we can only speculate on the origins of modern protein synthesis and the genetic code, several experimental observations have provided plausible scenarios.

In modern cells, some short peptides (such as antibiotics) are synthesized without the ribosome; peptide synthetase enzymes assemble these peptides, with their proper sequence of amino acids, without mRNAs to guide their synthesis. It is plausible that this non-coded, primitive version of protein synthesis first developed during the RNA world where it would have been catalyzed by RNA molecules. This idea presents no conceptual difficulties because, as we have seen, rRNA catalyzes peptide bond formation in present-day cells. We also know that ribozymes created in the laboratory can perform specific aminoacylation reactions; that is, they can match specific amino acids to specific tRNAs. It is therefore possible that tRNA-like adapters, each matched to a specific amino acid, could have arisen in the RNA world, marking the beginnings of a genetic code.

In principle, other RNAs (the precursors to mRNAs) could have served as crude templates to direct the nonrandom polymerization of a few different amino acids. Any RNA that helped guide the synthesis of a useful polypeptide would have a great advantage in the evolutionary struggle for survival. We can envision a relatively nonspecific peptidyl transferase ribozyme, which, over time, grew larger and acquired the ability to position charged tRNAs accurately on RNA templates—leading eventually to the modern ribosome. Once protein

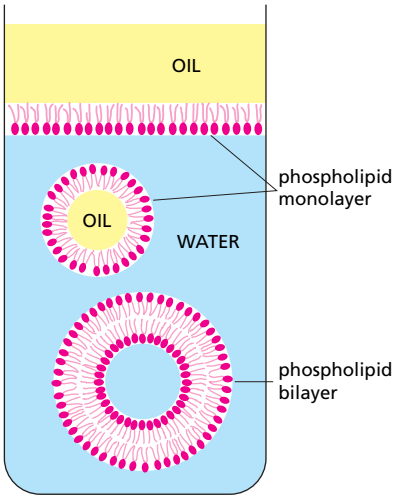


Figure 6–108 Formation of membrane by phospholipids. Because these molecules have hydrophilic heads and lipophilic tails, they align themselves at an oil/water interface with their heads in the water and their tails in the oil. In the water they associate to form closed bilayer vesicles in which the lipophilic tails are in contact with one another and the hydrophilic heads are exposed to the water.

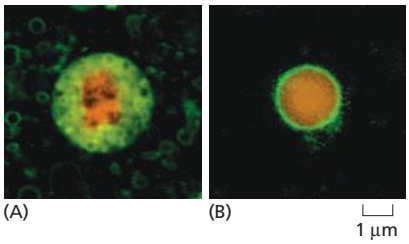


Figure 6–109 Encapsulation of RNA by simple amphiphilic molecules. For these experiments, the clay mineral montmorillonite was used to bring together RNA and fatty acids. (A) A montmorillonite particle, coated by RNA (*red*) has become trapped inside a fatty acid vesicle (*green*). (B) RNA *(red)* in solution has been encapsulated by fatty acids *(green)*. These experiments show that montmorillonite can greatly accelerate the spontaneous generation of vesicles from amphiphilic molecules and trap RNA inside them. It has been hypothesized that conceptually similar actions may have led to the first primitive cells on Earth. (From M.M. Hanczyc et al., *Science* 302:618–622, 2003. With permission from AAAS.)

synthesis evolved, the transition to a protein-dominated world could proceed, with proteins eventually taking over the majority of catalytic and structural tasks because of their greater versatility, with 20 rather than 4 different subunits. Although the scenarios just discussed are highly speculative, the known properties of RNA molecules are consistent with these ideas.

All Present-Day Cells Use DNA as Their Hereditary Material

If the evolutionary speculations embodied in the RNA world hypothesis are correct, early cells would have differed fundamentally from the cells we know today in having their hereditary information stored in RNA rather than in DNA (**Figure 6–110**). Evidence that RNA arose before DNA in evolution can be found in the chemical differences between them. Ribose, like glucose and other simple carbohydrates, can be formed from formaldehyde (HCHO), a simple chemical which is readily produced in laboratory experiments that attempt to simulate conditions on the primitive Earth. The sugar deoxyribose is harder to make, and in present-day cells it is produced from ribose in a reaction catalyzed by a protein enzyme, suggesting that ribose predates deoxyribose in cells. Presumably, DNA appeared on the scene later, but then proved more suitable than RNA as a permanent repository of genetic information. In particular, the deoxyribose in its sugar-phosphate backbone makes chains of DNA chemically more stable than chains of RNA, so that much greater lengths of DNA can be maintained without breakage.

The other differences between RNA and DNA—the double-helical structure of DNA and the use of thymine rather than uracil—further enhance DNA stability by making the many unavoidable accidents that occur to the molecule much easier to repair, as discussed in detail in Chapter 5 (see pp. 296–297 and 300–301).

Summary

From our knowledge of present-day organisms and the molecules they contain, it seems likely that the development of the directly autocatalytic mechanisms fundamental to living systems began with the evolution of families of molecules that could catalyze their own replication. With time, a family of cooperating RNA catalysts probably developed the ability to direct the synthesis of polypeptides. DNA is likely to have been a late addition: as the accumulation of additional protein catalysts allowed more efficient and complex cells to evolve, the DNA double helix replaced RNA as a more stable molecule for storing the increased amounts of genetic information required by such cells.

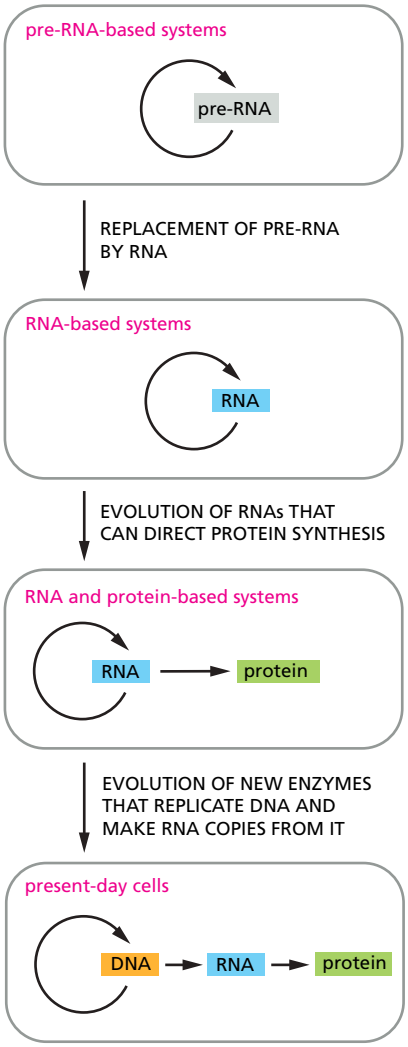


Figure 6–110 The hypothesis that RNA preceded DNA and proteins in evolution. In the earliest cells, pre-RNA molecules would have had combined genetic, structural, and catalytic functions and RNA would have gradually taken over these functions. In present-day cells, DNA is the repository of genetic information, and proteins perform the vast majority of catalytic functions in cells. RNA primarily functions today as a go-between in protein synthesis, although it remains a catalyst for a small number of crucial reactions.

PROBLEMS

Which statements are true? Explain why or why not.

6–1 The consequences of errors in transcription are less than those of errors in RNA replication.

6–2 Since introns are largely genetic "junk," they do not have to be removed precisely from the primary transcript during RNA splicing.

6–3 Wobble pairing occurs between the first position in the codon and the third position in the anticodon.

6–4 Protein enzymes are thought to greatly outnumber ribozymes in modern cells because they catalyze a much greater variety of reactions at much faster rates than ribozymes.

Discuss the following problems.

6–5 In which direction along the template must the RNA polymerase in **Figure Q6–1** be moving to have generated the supercoiled structures that are shown? Would you expect supercoils to be generated if the RNA polymerase were free to rotate about the axis of the DNA as it progressed along the template?

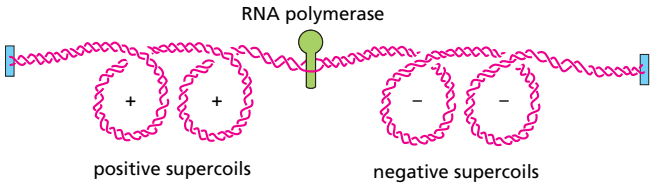


Figure Q6–1 Supercoils around a moving RNA polymerase (Problem 6–5).

6–6 Phosphates are attached to the CTD (C-terminal domain) of RNA polymerasse II during transcription. What are the various roles of RNA polymerase II CTD phosphorylation?

6–7 The human α-tropomyosin gene is alternatively spliced to produce several forms of α-tropomyosin mRNA in various cell types (**Figure Q6–2**). For all forms of the mRNA, the encoded protein sequence is the same for exons 1 and

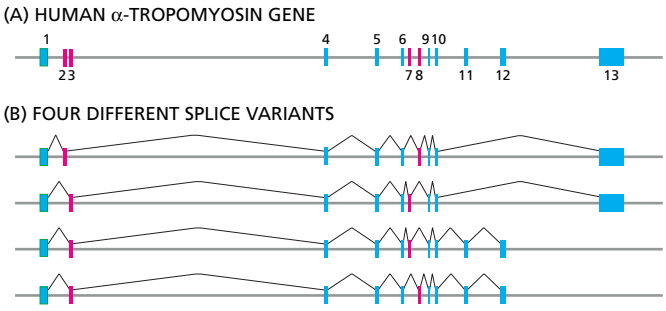


Figure Q6–2 Alternatively spliced mRNAs from the human α-tropomyosin gene (Problem 6–7). (A) Exons in the human α-tropomyosin gene. The locations and relative sizes of exons are shown by the *blue* and *red rectangles*. (B) Splicing patterns for four α-tropomyosin mRNAs. Splicing is indicated by *lines* connecting the exons that are included in the mRNA.

10. Exons 2 and 3 are alternative exons used in different mRNAs, as are exons 7 and 8. Which of the following statements about exons 2 and 3 is the most accurate? Is that statement also the most accurate one for exons 7 and 8? Explain your answers.

A. Exons 2 and 3 must have the same number of nucleotides.

B. Exons 2 and 3 must each contain an integral number of codons (that is, the number of nucleotides divided by 3 must be an integer).

C. Exons 2 and 3 must each contain a number of nucleotides that when divided by 3 leaves the same remainder (that is, 0, 1, or 2).

6–8 After treating cells with a chemical mutagen, you isolate two mutants. One carries alanine and the other carries methionine at a site in the protein that normally contains valine (**Figure Q6–3**). After treating these two mutants again with the mutagen, you isolate mutants from each that now carry threonine at the site of the original valine (Figure Q6–3). Assuming that all mutations involve single nucleotide changes, deduce the codons that are used for valine, methionine, threonine, and alanine at the affected site. Would you expect to be able to isolate valine-to-threonine mutants in one step?

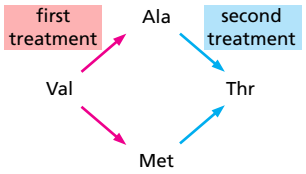


Figure Q6–3 Two rounds of mutagenesis and the altered amino acids at a single position in a protein (Problem 6–8).

6–9 The elongation factor EF-Tu introduces two short delays between codon–anticodon base-pairing and formation of the peptide bond. These delays increase the accuracy of protein synthesis. Describe these delays and explain how they improve the fidelity of translation.

6–10 Both Hsp60-like and Hsp70 molecular chaperones share an affinity for exposed hydrophobic patches on proteins, using them as indicators of incomplete folding. Why do you suppose hydrophobic patches serve as critical signals for the folding status of a protein?

6–11 Most proteins require molecular chaperones to assist in their correct folding. How do you suppose the chaperones themselves manage to fold correctly?

6–12 What is so special about RNA that makes it such an attractive evolutionary precursor to DNA and protein? What is it about DNA that makes it a better material than RNA for storage of genetic information?

6–13 If an RNA molecule could form a hairpin with a symmetric internal loop, as shown in **Figure Q6–4**, could the complement of this RNA form a similar structure? If so, would there be any regions of the two structures that are identical? Which ones?

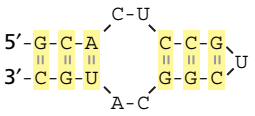


Figure Q6–4 An RNA hairpin with a symmetric internal loop (Problem 6–13).

REFERENCES

General

Berg JM, Tymoczko JL & Stryer L (2006) Biochemistry, 6th ed. New York: WH Freeman.
Brown TA (2002) Genomes 2, 2nd ed. New York: Wiley-Liss.
Gesteland RF, Cech TR & Atkins JF (eds) (2006) The RNA World, 3rd ed. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.
Hartwell L, Hood L, Goldberg ML et al (2006) Genetics: from Genes to Genomes, 3rd ed. Boston: McGraw Hill.
Lodish H, Berk A, Kaiser C et al (2007) Molecular Cell Biology, 6th ed. New York: WH Freeman.
Stent GS (1971) Molecular Genetics: An Introductory Narrative. San Francisco: WH Freeman.
The Genetic Code (1966) *Cold Spring Harb. Symp Quant Biol* 31.
The Ribosome (2001) *Cold Spring Harb Symp Quant Biol* 66.
Watson JD, Baker TA, Bell SP et al (2003) Molecular Biology of the Gene, 5th ed. Menlo Park, CA: Benjamin Cummings.

From DNA to RNA

Bentley DL (2005) Rules of engagement: co-transcriptional recruitment of pre-mRNA processing factors. *Curr Opin Cell Biol* 17:251–256.
Berget SM, Moore C & Sharp PA (1977) Spliced segments at the 5′ terminus of adenovirus 2 late mRNA. *Proc Natl Acad Sci USA* 74:3171–3175.
Black DL (2003). Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem* 72:291–336.
Brenner S, Jacob F & Meselson M (1961) An unstable intermediate carrying information from genes to ribosomes for protein synthesis. *Nature* 190:576–581.
Cate JH, Gooding AR, Podell E et al. (1996) Crystal structure of a group I ribozyme domain: principles of RNA packing. *Science* 273:1678–1685.
Chow LT, Gelinas RE, Broker TR et al (1977) An amazing sequence arrangement at the 5′ ends of adenovirus 2 messenger RNA. *Cell* 12:1–8.
Cramer P (2002) Multisubunit RNA polymerases. *Curr Opin Struct Biol* 12:89–97.
Daneholt B (1997) A look at messenger RNP moving through the nuclear pore. *Cell* 88:585–588.
Dreyfuss G, Kim VN, & Kataoka N (2002) Messenger-RNA-binding proteins and the messages they carry. *Nature Rev Mol Cell Biol* 3:195–205.
Houseley J, LaCava J & Tollervey D (2006) RNA-quality control by the exosome. *Nature Rev Mol Cell Biol* 7:529–539.
Izquierdo JM, & Valcárcel J (2006) A simple principle to explain the evolution of pre-mRNA splicing. *Genes Dev* 20:1679–1684.
Kornberg RD (2005) Mediator and the mechanism of transcriptional activation. *Trends Biochem Sci* 30:235–239.
Malik S & Roeder RG (2005) Dynamic regulation of pol II transcription by the mammalian Mediator complex. *Trends Biochem Sci* 30:256–263.
Matsui T, Segall J, Weil PA & Roeder RG (1980) Multiple factors required for accurate initiation of transcription by purified RNA polymerase II. *J Biol Chem* 255:11992–11996.
Patel AA & Steitz JA (2003) Splicing double: insights from the second spliceosome. *Nature Rev Mol Cell Biol* 4|:960–970.
Phatnani HP & Greenleaf AL (2006) Phosphorylation and functions of the RNA polymerase II CTD. *Genes Dev* 20:2922–2936.
Query CC & Konarska MM (2006) Splicing fidelity revisited. *Nature Struct Mol Biol* 13:472–474.
Ruskin B, Krainer AR, Maniatis T et al (1984) Excision of an intact intron as a novel lariat structure during pre-mRNA splicing *in vitro*. *Cell* 38:317–331.
Spector DL (2003) The dynamics of chromosome organization and gene regulation. *Annu Rev Biochem* 72:573–608.
Staley JP & Guthrie C (1998) Mechanical devices of the spliceosome: motors, clocks, springs, and things. *Cell* 92:315–326.
Thomas MC & Chiang CM (2006) The general transcription machinery and general cofactors. *Critical Rev Biochem Mol Biol* 41:105–178.
Wang D, Bushnell DA, Westover KD et al (2006) Structural basis of transcription: role of the trigger loop in substrate specificity and catalysis. *Cell* 127:941–954.

From RNA to Protein

Allen GS & Frank J (2007) Structural insights on the translation initiation complex: ghosts of a universal initiation complex. *Mol Microbiol* 63:941–950.
Anfinsen CB (1973) Principles that govern the folding of protein chains. *Science* 181:223–230.
Brunelle JL, Youngman EM, Sharma D et al (2006) The interaction between C75 of tRNA and the A loop of the ribosome stimulates peptidyl transferase activity. *RNA* 12:33–39.
Chien P, Weissman JS, & DePace AH (2004). Emerging principles of conformation-based prion inheritance. *Annu Rev Biochem* 73:617–656.
Crick FHC (1966) The genetic code: III. *Sci Am* 215:55–62.
Hershko A, Ciechanover A & Varshavsky A (2000) The ubiquitin system. *Nature Med* 6:1073–1081.
Ibba M & Soll D (2000) Aminoacyl-tRNA synthesis. *Annu Rev Biochem* 69:617–650.
Kozak M (1992) Regulation of translation in eukaryotic systems. *Annu Rev Cell Biol* 8:197–225.
Kuzmiak HA, & Maquat LE (2006) Applying nonsense-mediated mRNA decay research to the clinic: progress and challenges. *Trends Mol Med* 12:306–316.
Moore PB & Steitz TA (2005) The ribosome revealed. *Trends Biochem Sci* 30:281–283.
Noller HF (2005) RNA structure: reading the ribosome. *Science* 309:1508–1514.
Ogle JM, Carter AP & Ramakrishnan V (2003) Insights into the decoding mechanism from recent ribosome structures. *Trends Biochem Sci* 28:259–266.
Prusiner SB (1998) Nobel lecture. Prions. *Proc Natl Acad Sci USA* 95:13363–13383.
Rehwinkel J, Raes J & Izaurralde E (2006) Nonsense-mediated mRNA decay: Target genes and functional diversification of effectors. *Trends Biochem Sci* 31:639–646.
Sauer RT, Bolon DN, Burton BM et al (2004) Sculpting the proteome with AAA(+) proteases and disassembly machines. *Cell* 119:9–18.
Shorter J & Lindquist S (2005) Prions as adaptive conduits of memory and inheritance. *Nature Rev Genet* 6:435–450.
Varshavsky A (2005) Regulated protein degradation. *Trends in Biochem Sci* 30:283–286.
Voges D, Zwickl P & Baumeister W (1999) The 26S proteasome: a molecular machine designed for controlled proteolysis. *Annu Rev Biochem* 68:1015–1068.
Weissmann C (2005) Birth of a prion: spontaneous generation revisited. *Cell* 122:165–168.
Young JC, Agashe VR, Siegers K et al (2004) Pathways of chaperone-mediated protein folding in the cytosol. *Nature Rev Mol Cell Biol* 5:781–791.

The RNA World and the Origins of Life

Joyce GF (1992) Directed molecular evolution. *Sci Am* 267:90–97.
Orgel L (2000) Origin of life. A simpler nucleic acid. *Science* 290:1306–1307.
Kruger K, Grabowski P, Zaug P et al (1982) Self-splicing RNA: Autoexcision and autocyclization of the ribosomal RNA intervening seuence of Tetrahymena. *Cell* 31:147–157.
Silverman SK (2003) Rube Goldberg goes (ribo)nuclear? Molecular switches and sensors made from RNA. *RNA* 9:377–383.
Szostak JW, Bartel DP & Luisi PL (2001) Synthesizing life. *Nature* 409:387–390.

7

Control of Gene Expression

An organism's DNA encodes all of the RNA and protein molecules required to construct its cells. Yet a complete description of the DNA sequence of an organism—be it the few million nucleotides of a bacterium or the few billion nucleotides of a human—no more enables us to reconstruct the organism than a list of English words enables us to reconstruct a play by Shakespeare. In both cases, the problem is to know how the elements in the DNA sequence or the words on the list are used. Under what conditions is each gene product made, and, once made, what does it do?

In this chapter we discuss the first half of this problem—the rules and mechanisms by which a subset of the genes is selectively expressed in each cell. The mechanisms that control the expression of genes operate at many levels, and we discuss the different levels in turn. We begin with an overview of some basic principles of gene control in multicellular organisms.

In This Chapter

AN OVERVIEW OF GENE CONTROL 411

DNA-BINDING MOTIFS IN GENE REGULATORY PROTEINS 416

HOW GENETIC SWITCHES WORK 432

THE MOLECULAR GENETIC MECHANISMS THAT CREATE SPECIALIZED CELL TYPES 454

POST-TRANSCRIPTIONAL CONTROLS 477

AN OVERVIEW OF GENE CONTROL

The different cell types in a multicellular organism differ dramatically in both structure and function. If we compare a mammalian neuron with a lymphocyte, for example, the differences are so extreme that it is difficult to imagine that the two cells contain the same genome (**Figure 7–1**). For this reason, and because cell differentiation is often irreversible, biologists originally suspected that genes might be selectively lost when a cell differentiates. We now know, however, that cell differentiation generally depends on changes in gene expression rather than on any changes in the nucleotide sequence of the cell's genome.

The Different Cell Types of a Multicellular Organism Contain the Same DNA

The cell types in a multicellular organism become different from one another because they synthesize and accumulate different sets of RNA and protein molecules. Evidence that they generally do this without altering the sequence of their DNA comes from a classic set of experiments in frogs. When the nucleus of a fully differentiated frog cell is injected into a frog egg whose nucleus has been removed, the injected donor nucleus is capable of directing the recipient egg to produce a normal tadpole (**Figure 7–2**A). Because the tadpole contains a full range of differentiated cells that derived their DNA sequences from the nucleus of the original donor cell, it follows that the differentiated donor cell cannot have lost any important DNA sequences. A similar conclusion has been reached in experiments performed with various plants. Here differentiated pieces of plant tissue are placed in culture and then dissociated into single cells. Often, one of these individual cells can regenerate an entire adult plant (Figure 7–2B). Finally, this same principle has been demonstrated in mammals, including sheep, cattle, pigs, goats, dogs, and mice by introducing nuclei from somatic cells into enucleated eggs; when placed into surrogate mothers, some of these eggs (called reconstructed zygotes) develop into healthy animals (Figure 7–2C).

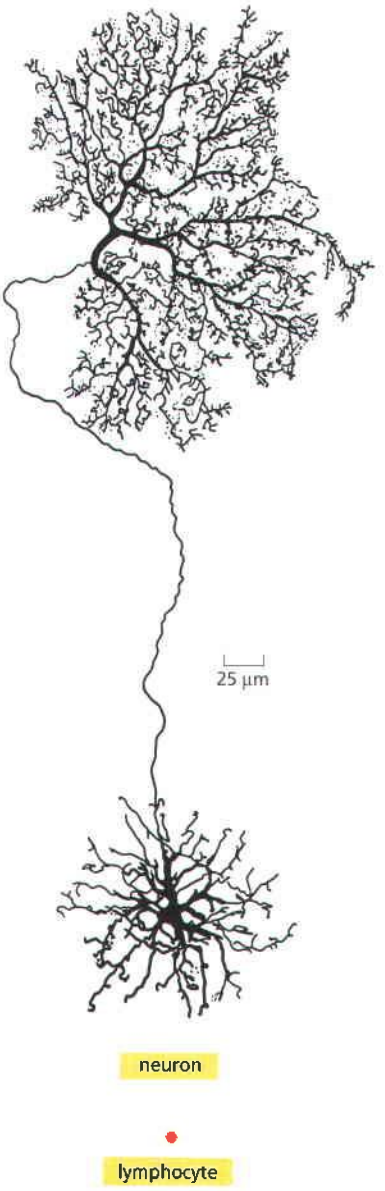


Figure 7–1 A mammalian neuron and a lymphocyte. The long branches of this neuron from the retina enable it to receive electrical signals from many cells and carry those signals to many neighboring cells. The lymphocyte is a white blood cell involved in the immune response to infection and moves freely through the body. Both of these cells contain the same genome, but they express different RNAs and proteins. (From B.B. Boycott, Essays on the Nervous System [R. Bellairs and E.G. Gray, eds.]. Oxford, UK: Clarendon Press, 1974.)

Further evidence that large blocks of DNA are not lost or rearranged during vertebrate development comes from comparing the detailed banding patterns detectable in condensed chromosomes at mitosis (see Figure 4–11). By this criterion the chromosome sets of differentiated cells in the human body appear to be identical. Moreover, comparisons of the genomes of different cells based on recombinant DNA technology have confirmed, as a general rule, that the changes in gene expression that underlie the development of multicellular organisms do not rely on changes in the DNA sequences of the corresponding genes. There are, however, a few cases where DNA rearrangements of the genome take place during the development of an organism—most notably, in generating the diversity of the immune system of mammals, which we discuss in Chapter 25.

Different Cell Types Synthesize Different Sets of Proteins

As a first step in understanding cell differentiation, we would like to know how many differences there are between any one cell type and another. Although we still do not have a detailed answer to this fundamental question, we can make certain general statements.

1. Many processes are common to all cells, and any two cells in a single organism therefore have many proteins in common. These include the structural proteins of chromosomes, RNA polymerases, DNA repair enzymes, ribosomal proteins, enzymes involved in the central reactions of metabolism, and many of the proteins that form the cytoskeleton.
2. Some proteins are abundant in the specialized cells in which they function and cannot be detected elsewhere, even by sensitive tests. Hemoglobin, for example, can be detected only in red blood cells.
3. Studies of the number of different mRNAs suggest that, at any one time, a typical human cell expresses 30–60% of its approximately 25,000 genes. When the patterns of mRNAs in a series of different human cell lines are compared, it is found that the level of expression of almost every active gene varies from one cell type to another. A few of these differences are striking, like that of hemoglobin noted above, but most are much more subtle. Even genes that are expressed in all cell types vary in their level of expression from one cell type to the next. The patterns of mRNA abundance (determined using DNA microarrays, discussed in Chapter 8) are so characteristic of cell type that they can be used to type human cancer cells of uncertain tissue origin (**Figure 7–3**).
4. Although the differences in mRNAs among specialized cell types are striking, they nonetheless underestimate the full range of differences in the pattern of protein production. As we shall see in this chapter, there are many steps after transcription at which gene expression can be regulated. For example, alternative splicing can produce a whole family of proteins from a single gene. Finally, proteins can be covalently modified after they are synthesized. Therefore a better way of appreciating the radical differences in gene expression between cell types is through methods that directly display the levels of proteins and their post-translational modifications (**Figure 7–4**).

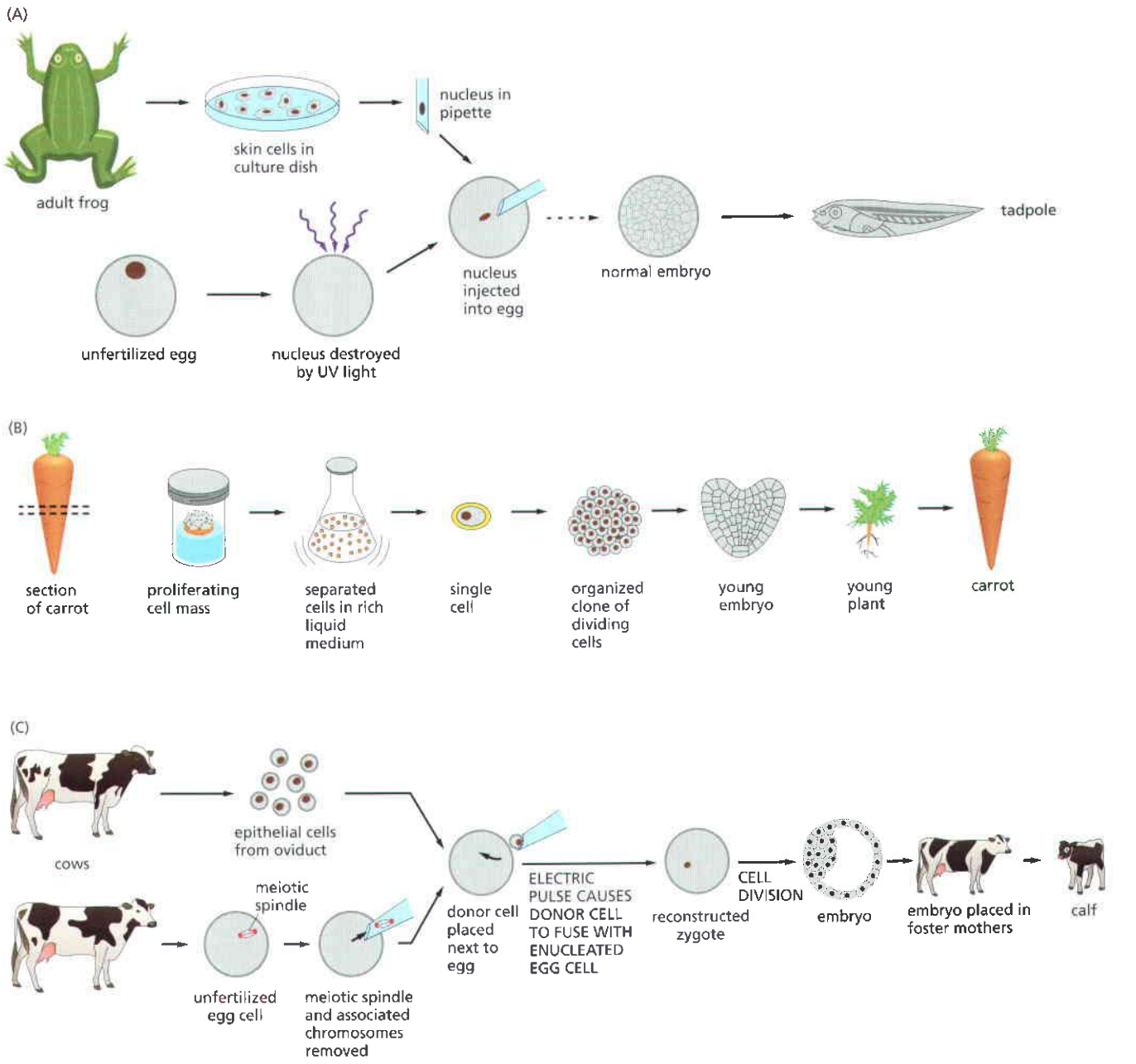


Figure 7–2 Evidence that a differentiated cell contains all the genetic instructions necessary to direct the formation of a complete organism. (A) The nucleus of a skin cell from an adult frog transplanted into an enucleated egg can give rise to an entire tadpole. The *broken arrow* indicates that, to give the transplanted genome time to adjust to an embryonic environment, a further transfer step is required in which one of the nuclei is taken from the early embryo that begins to develop and is put back into a second enucleated egg. (B) In many types of plants, differentiated cells retain the ability to "dedifferentiate," so that a single cell can form a clone of progeny cells that later give rise to an entire plant. (C) A differentiated cell nucleus from an adult cow introduced into an enucleated egg from a different cow can give rise to a calf. Different calves produced from the same differentiated cell donor are genetically identical and are therefore clones of one another. (A, modified from J.B. Gurdon, *Sci. Am.* 219:24–35, 1968. With permission from Scientific American.)

External Signals Can Cause a Cell to Change the Expression of Its Genes

Most of the specialized cells in a multicellular organism are capable of altering their patterns of gene expression in response to extracellular cues. If a liver cell is exposed to a glucocorticoid hormone, for example, the production of several specific proteins is dramatically increased. Glucocorticoids are released in the body during periods of starvation or intense exercise and signal the liver to increase the

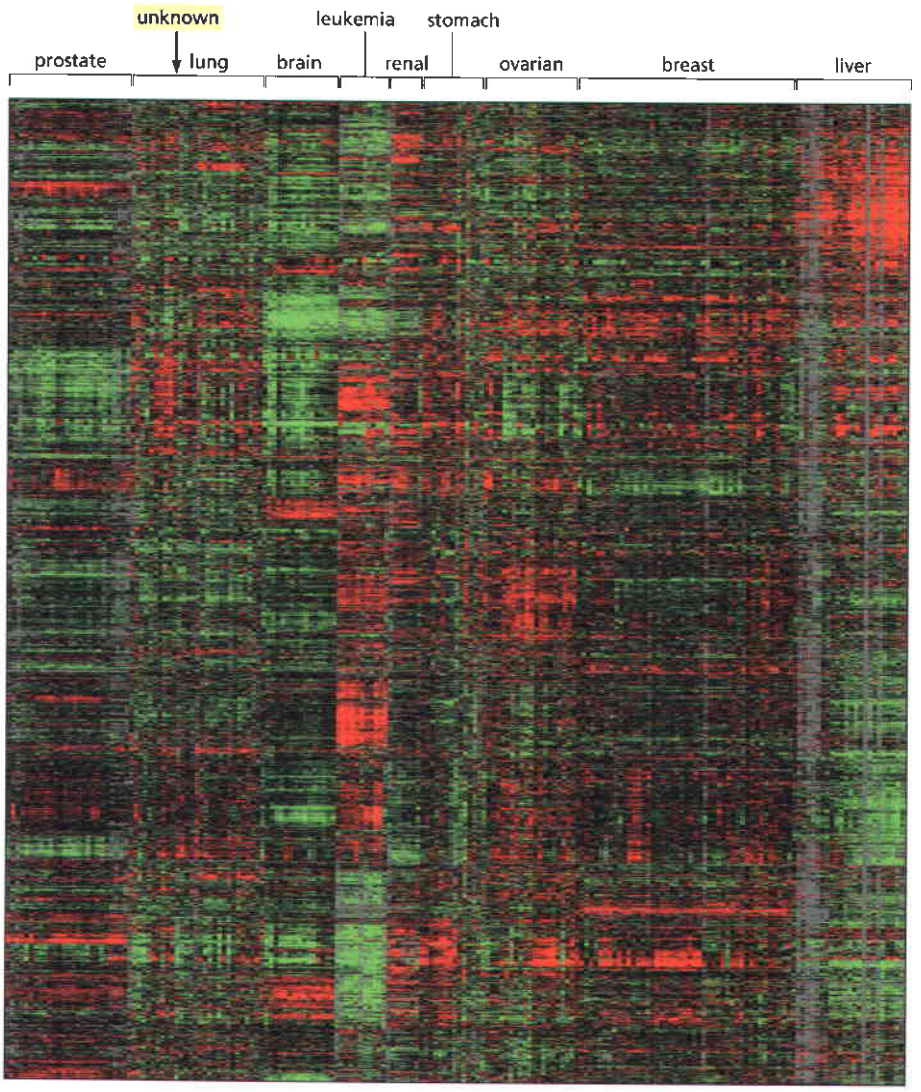


Figure 7–3 Differences in mRNA expression patterns among different types of human cancer cells. This figure summarizes a very large set of measurements in which the mRNA levels of 1800 selected genes (arranged *top* to *bottom*) were determined for 142 different human tumor cell lines (arranged *left* to *right*), each from a different patient. Each small *red* bar indicates that the given gene in the given tumor is transcribed at a level significantly higher than the average across all the cell lines. Each small *green* bar indicates a less-than-average expression level, and each *black* bar denotes an expression level that is close to average across the different tumors. The procedure used to generate these data—mRNA isolation followed by hybridization to DNA microarrays—is described in Chapter 8 (pp. 574–575). The figure shows that the relative expression levels of each of the 1800 genes analyzed vary among the different tumors (seen by following a given gene from *left* to *right* across the figure). This analysis also shows that each type of tumor has a characteristic gene expression pattern. This information can be used to "type" cancer cells of unknown tissue origin by matching the gene expression profiles to those of known tumors. For example, the unknown sample in the figure has been identified as a lung cancer. (Courtesy of Patrick O. Brown, David Botstein, and the Stanford Expression Collaboration.)

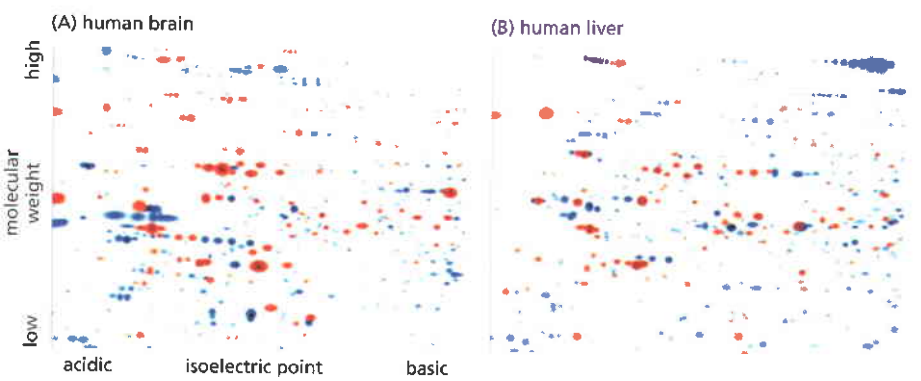


Figure 7–4 Differences in the proteins expressed by two human tissues. In each panel, the proteins are displayed using two-dimensional polyacrylamide-gel electrophoresis (see pp. 521–522). The proteins have been separated by molecular weight (*top* to *bottom*) and isoelectric point, the pH at which the protein has no net charge (*right* to *left*). The protein spots artificially colored *red* are common to both samples; those in *blue* are specific to one of the two tissues. The differences between the two tissue samples vastly outweigh their similarities: even for proteins that are shared between the two tissues, their relative abundances are usually different. Note that this technique separates proteins by both size and charge; therefore a protein that has, for example, several different phosphorylation states will appear as a series of *horizontal spots* (see *upper right-hand* portion of *right* panel). Only a small portion of the complete protein spectrum is shown for each sample. Although two-dimensional gel electrophoresis provides a simple way to visualize the differences between two protein samples, methods based on mass spectrometry (see pp. 519–521) provide much more detailed information and are therefore more commonly used. (Courtesy of Tim Myers and Leigh Anderson, Large Scale Biology Corporation.)

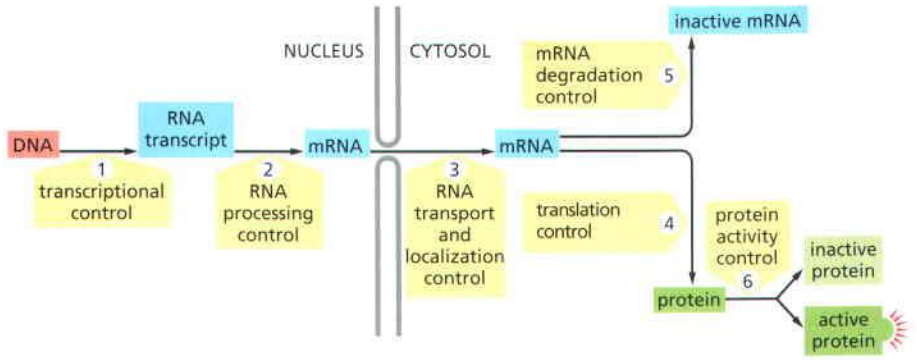


Figure 7–5 Six steps at which eucaryotic gene expression can be controlled. Controls that operate at steps 1 through 5 are discussed in this chapter. Step 6, the regulation of protein activity, occurs largely through covalent post-translational modifications including phosphorylation, acetylation, and ubiquitylation (see Table 3–3, p. 186) and is discussed in many chapters throughout the book.

production of glucose from amino acids and other small molecules; the set of proteins whose production is induced includes enzymes such as tyrosine aminotransferase, which helps to convert tyrosine to glucose. When the hormone is no longer present, the production of these proteins drops to its normal level.

Other cell types respond to glucocorticoids differently. Fat cells, for example, reduce the production of tyrosine aminotransferase, while some other cell types do not respond to glucocorticoids at all. These examples illustrate a general feature of cell specialization: different cell types often respond differently to the same extracellular signal. Underlying such adjustments that occur in response to extracellular signals, there are features of the gene expression pattern that do not change and give each cell type its permanently distinctive character.

Gene Expression Can Be Regulated at Many of the Steps in the Pathway from DNA to RNA to Protein

If differences among the various cell types of an organism depend on the particular genes that the cells express, at what level is the control of gene expression exercised? As we saw in the previous chapter, there are many steps in the pathway leading from DNA to protein. We now know that all of them can in principle be regulated. Thus a cell can control the proteins it makes by (1) controlling when and how often a given gene is transcribed (**transcriptional control**), (2) controlling the splicing and processing of RNA transcripts (**RNA processing control**), (3) selecting which completed mRNAs are exported from the nucleus to the cytosol and determining where in the cytosol they are localized (**RNA transport and localization control**), (4) selecting which mRNAs in the cytoplasm are translated by ribosomes (**translational control**), (5) selectively destabilizing certain mRNA molecules in the cytoplasm (**mRNA degradation control**), or (6) selectively activating, inactivating, degrading, or locating specific protein molecules after they have been made (**protein activity control**) (Figure 7–5).

For most genes transcriptional controls are paramount. This makes sense because, of all the possible control points illustrated in Figure 7–5, only transcriptional control ensures that the cell will not synthesize superfluous intermediates. In the following sections we discuss the DNA and protein components that perform this function by regulating the initiation of gene transcription. We shall return at the end of the chapter to the many additional ways of regulating gene expression.

Summary

The genome of a cell contains in its DNA sequence the information to make many thousands of different protein and RNA molecules. A cell typically expresses only a fraction of its genes, and the different types of cells in multicellular organisms arise because different sets of genes are expressed. Moreover, cells can change the pattern of genes they express in response to changes in their environment, such as signals from other cells. Although all of the steps involved in expressing a gene can in principle be regulated, for most genes the initiation of RNA transcription is the most important point of control.

DNA-BINDING MOTIFS IN GENE REGULATORY PROTEINS

How does a cell determine which of its thousands of genes to transcribe? As outlined in Chapter 6, the transcription of each gene is controlled by a regulatory region of DNA relatively near the site where transcription begins. Some regulatory regions are simple and act as switches thrown by a single signal. Many others are complex and resemble tiny microprocessors, responding to a variety of signals that they interpret and integrate in order to switch their neighboring gene on or off. Whether complex or simple, these switching devices are found in all cells and are composed of two types of fundamental components: (1) short stretches of DNA of defined sequence and (2) *gene regulatory proteins* that recognize and bind to this DNA.

We begin our discussion of gene regulatory proteins by describing how they were discovered.

Gene Regulatory Proteins Were Discovered Using Bacterial Genetics

Genetic analyses in bacteria carried out in the 1950s provided the first evidence for the existence of **gene regulatory proteins** (often loosely called "transcription factors") that turn specific sets of genes on or off. One of these regulators, the *lambda repressor*, is encoded by a bacterial virus, *bacteriophage lambda*. The repressor shuts off the viral genes that code for the protein components of new virus particles and thereby enables the viral genome to remain a silent passenger in the bacterial chromosome, multiplying with the bacterium when conditions are favorable for bacterial growth (see Figure 5–78). The lambda repressor was among the first gene regulatory proteins to be characterized, and it remains one of the best understood, as we discuss later. Other bacterial regulators respond to nutritional conditions by shutting off genes encoding specific sets of metabolic enzymes when they are not needed. The *Lac repressor*, the first of these bacterial proteins to be recognized, turns off the production of the proteins responsible for lactose metabolism when this sugar is absent from the medium.

The first step toward understanding gene regulation was the isolation of mutant strains of bacteria and bacteriophage lambda that were unable to shut off specific sets of genes. It was proposed at the time, and later proven, that most of these mutants were deficient in proteins acting as specific repressors for these sets of genes. Because these proteins, like most gene regulatory proteins, are present in small quantities, it was difficult and time-consuming to isolate them. They were eventually purified by fractionating cell extracts. Once isolated, the proteins were shown to bind to specific DNA sequences close to the genes that they regulate. The precise DNA sequences that they recognized were then determined by a combination of classical genetics and methods for studying protein–DNA interactions discussed later in this chapter.

The Outside of the DNA Helix Can Be Read by Proteins

As discussed in Chapter 4, the DNA in a chromosome consists of a very long double helix (**Figure 7–6**). Gene regulatory proteins must recognize specific nucleotide sequences embedded within this structure. It was originally thought that these proteins might require direct access to the hydrogen bonds between base pairs in the interior of the double helix to distinguish between one DNA

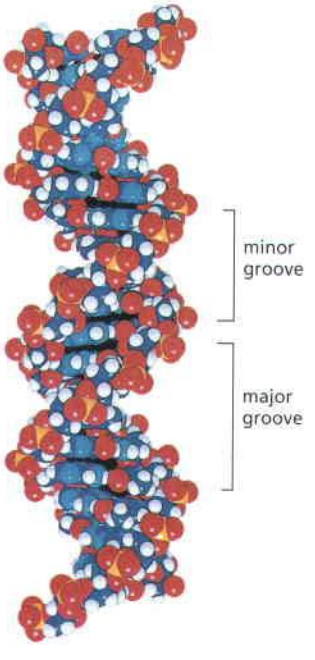


Figure 7–6 Double-helical structure of DNA. A space-filling model of DNA showing the major and minor grooves on the outside of the double helix. The atoms are colored as follows: carbon, *dark blue;* nitrogen, *light blue;* hydrogen, *white;* oxygen, *red;* phosphorus, *yellow.*

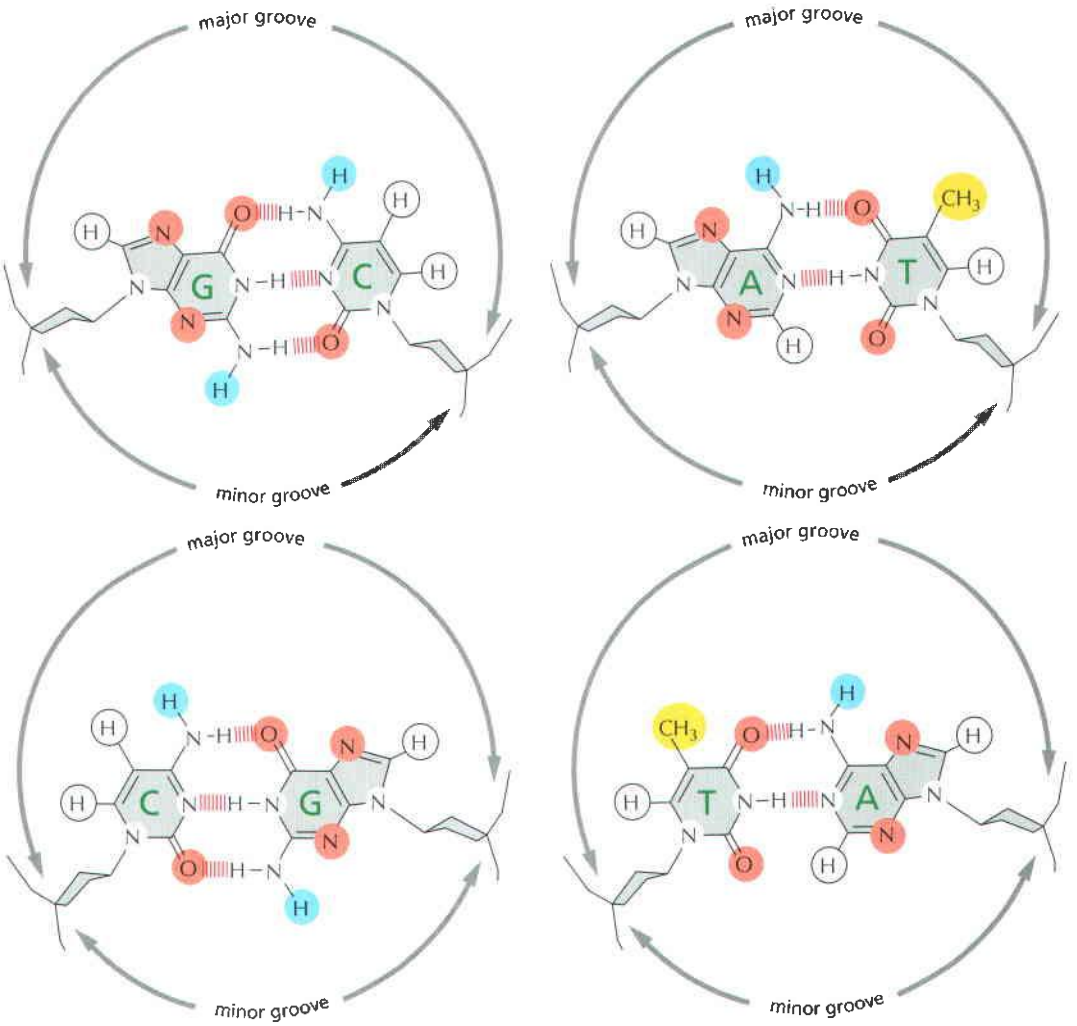


Figure 7–7 How the different base pairs in DNA can be recognized from their edges without the need to open the double helix. The four possible configurations of base pairs are shown, with potential hydrogen bond donors indicated in *blue*, potential hydrogen bond acceptors in *red*, and hydrogen bonds of the base pairs themselves as a series of short parallel *red* lines. Methyl groups, which form hydrophobic protuberances, are shown in *yellow*, and hydrogen atoms that are attached to carbons, and are therefore unavailable for hydrogen bonding, are *white*. (From C. Branden and J. Tooze, Introduction to Protein Structure, 2nd ed. New York: Garland Publishing, 1999.)

sequence and another. It is now clear, however, that the outside of the double helix is studded with DNA sequence information that gene regulatory proteins can recognize without having to open the double helix. The edge of each base pair is exposed at the surface of the double helix, presenting a distinctive pattern of hydrogen bond donors, hydrogen bond acceptors, and hydrophobic patches for proteins to recognize in both the major and minor groove (**Figure 7–7**). But only in the major groove are the patterns markedly different for each of the four base-pair arrangements (**Figure 7–8**). For this reason, gene regulatory proteins generally make specific contacts with the major groove—as we shall see.

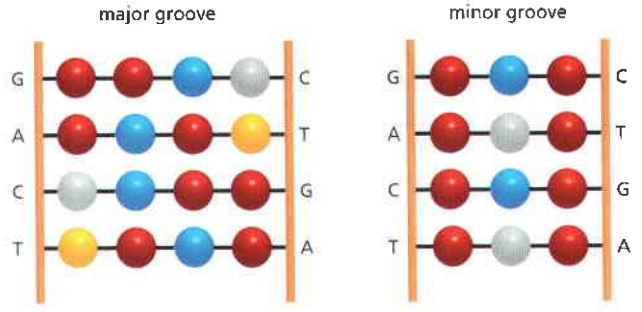


Figure 7–8 A DNA recognition code. The edge of each base pair, seen here looking directly at the major or minor groove, contains a distinctive pattern of hydrogen bond donors, hydrogen bond acceptors, and methyl groups. From the major groove, each of the four base-pair configurations projects a unique pattern of features. From the minor groove, however, the patterns are similar for G–C and C–G as well as for A–T and T–A. The color code is the same as that in Figure 7–7. (From C. Branden and J. Tooze, Introduction to Protein Structure, 2nd ed. New York: Garland Publishing, 1999.)

Short DNA Sequences Are Fundamental Components of Genetic Switches

A specific nucleotide sequence can be "read" as a pattern of molecular features on the surface of the DNA double helix. Particular nucleotide sequences, each typically less than 20 nucleotide pairs in length, function as fundamental components of genetic switches by serving as recognition sites for the binding of specific gene regulatory proteins. Thousands of such DNA sequences have been identified, each recognized by a different gene regulatory protein (or by a set of related gene regulatory proteins). Some of the gene regulatory proteins that are discussed in the course of this chapter are listed in Table 7–1, along with the DNA sequences that they recognize.

We now turn to the gene regulatory proteins themselves, the second fundamental component of genetic switches. We begin with the structural features that allow these proteins to recognize short, specific DNA sequences contained in a much longer double helix.

Gene Regulatory Proteins Contain Structural Motifs That Can Read DNA Sequences

Molecular recognition in biology generally relies on an exact fit between the surfaces of two molecules, and the study of gene regulatory proteins has provided some of the clearest examples of this principle. A gene regulatory protein recognizes a specific DNA sequence because the surface of the protein is extensively

Table 7–1 Some Gene Regulatory Proteins and the DNA Sequences That They Recognize

	NAME	DNA SEQUENCE RECOGNIZED*
Bacteria	Lac repressor	5′ AATTGTGAGCGGATAACAATT 3′ TTAACACTCGCCTATTGTTAA
	CAP	TGTGAGTTAGCTCACT ACACTCAATCGAGTGA
	Lambda repressor	TATCACCGCCAGAGGT ATAGTGGCGGTCTCCAT
Yeast	Gal4	CGGAGGACTGTCCTCCG GCCTCCTGACAGGAGGC
	Matα2	CATGTAATT GTACATTAA
	Gcn4	ATGACTCAT TACTGAGTA
Drosophila	Kruppel	AACGGGTTAA TTGCCCAATT
	Bicoid	GGGATTAGA CCCTAATCT
Mammals	Sp1	GGGCGG CCCGCC
	Oct1 Pou domain	ATGCAAAT TACGTTTA
	GATA1	TGATAG ACTATC
	MyoD	CAAATG GTTTAC
	p53	GGGCAAGTCT CCCGTTCAGA

*For convenience, only one recognition sequence, rather than a consensus sequence (see Figure 6–12), is given for each protein.

complementary to the special surface features of the double helix in that region. In most cases the protein makes a series of contacts with the DNA, involving hydrogen bonds, ionic bonds, and hydrophobic interactions. Although each individual contact is weak, the 20 or so that are typically formed at the protein–DNA interface add together to ensure that the interaction is both highly specific and very strong (**Figure 7–9**). In fact, DNA–protein interactions include some of the tightest and most specific molecular interactions known in biology.

Although each example of protein–DNA recognition is unique in detail, x-ray crystallographic and NMR spectroscopic studies of several hundred gene regulatory proteins have revealed that many of them contain one or another of a small set of DNA-binding structural motifs. These motifs generally use either α helices or β sheets to bind to the major groove of DNA; this groove, as we have seen, contains sufficient information to distinguish one DNA sequence from any other. The fit is so good that it has been suggested that the dimensions of the basic structural units of nucleic acids and proteins evolved together to permit these molecules to interlock.

The Helix–Turn–Helix Motif Is One of the Simplest and Most Common DNA-Binding Motifs

The first DNA-binding protein motif to be recognized was the **helix–turn–helix**. Originally identified in bacterial proteins, this motif has since been found in many hundreds of DNA-binding proteins from both eucaryotes and procaryotes. It is constructed from two α helices connected by a short extended chain of amino acids, which constitutes the "turn" (**Figure 7–10**). The two helices are held at a fixed angle, primarily through interactions between the two helices. The more C-terminal helix is called the *recognition helix* because it fits into the major groove of DNA; its amino acid side chains, which differ from protein to protein, play an important part in recognizing the specific DNA sequence to which the protein binds.

Outside the helix–turn–helix region, the structure of the various proteins that contain this motif can vary enormously (**Figure 7–11**). Thus each protein "presents" its helix–turn–helix motif to the DNA in a unique way, a feature thought to enhance the versatility of the helix–turn–helix motif by increasing the number of DNA sequences that the motif can be used to recognize. Moreover, in most of these proteins, parts of the polypeptide chain outside the helix–turn–helix domain also make important contacts with the DNA, helping to fine-tune the interaction.

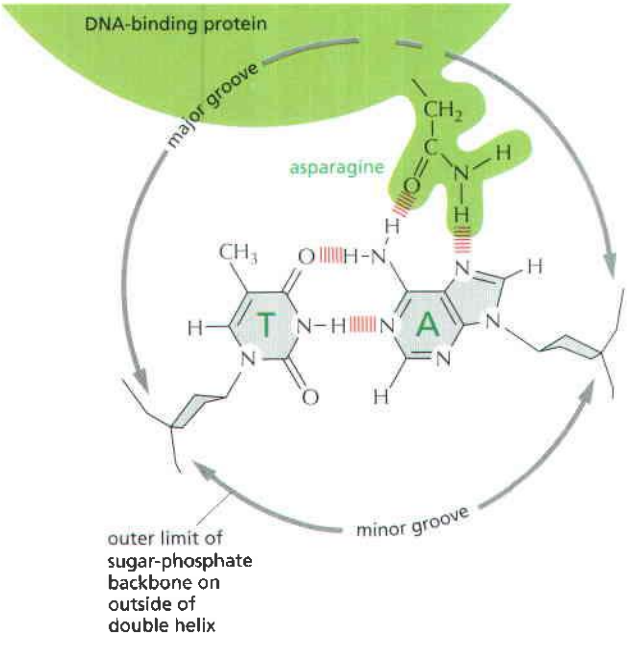


Figure 7–9 The binding of a gene regulatory protein to the major groove of DNA. Only a single contact is shown. Typically, the protein–DNA interface would consist of 10–20 such contacts, involving different amino acids, each contributing to the strength of the protein–DNA interaction.

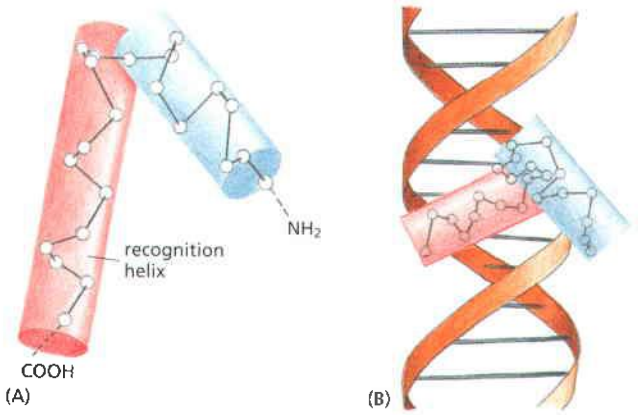


Figure 7–10 The DNA-binding helix–turn–helix motif. The motif is shown in (A), where each *white* circle denotes the central carbon of an amino acid. The C-terminal α helix *(red)* is called the recognition helix because it participates in sequence-specific recognition of DNA. As shown in (B), this helix fits into the major groove of DNA, where it contacts the edges of the base pairs (see also Figure 7–7). The N-terminal α-helix *(blue)* functions primarily as a structural component that helps to position the recognition helix.

The group of helix–turn–helix proteins shown in Figure 7–11 demonstrates a common feature of many sequence-specific DNA-binding proteins. They bind as symmetric dimers to DNA sequences that are composed of two very similar "half-sites," which are also arranged symmetrically (**Figure 7–12**). This arrangement allows each protein monomer to make a nearly identical set of contacts and enormously increases the binding affinity: as a first approximation, doubling the number of contacts doubles the free energy of the interaction and thereby *squares* the affinity constant.

Homeodomain Proteins Constitute a Special Class of Helix–Turn–Helix Proteins

Not long after the first gene regulatory proteins were discovered in bacteria, genetic analyses in the fruit fly *Drosophila* led to the characterization of an important class of genes, the *homeotic selector genes*, that play a critical part in orchestrating fly development. As discussed in Chapter 22, they have since proved to have a fundamental role in the development of higher animals as well. Mutations in these genes can cause one body part in the fly to be converted into another, showing that the proteins they encode control critical developmental decisions.

When the nucleotide sequences of several homeotic selector genes were determined in the early 1980s, each proved to code for an almost identical stretch of 60 amino acids that defines this class of proteins and is termed the **homeodomain**. When the three-dimensional structure of the homeodomain was determined, it was seen to contain a helix–turn–helix motif related to that of

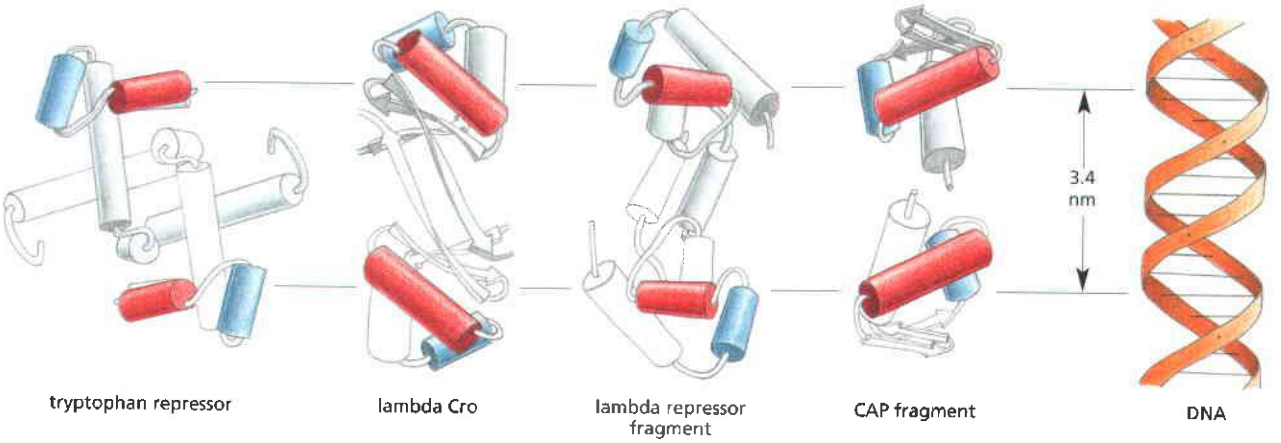


Figure 7–11 Some helix–turn–helix DNA-binding proteins. All of the proteins bind DNA as dimers in which the two copies of the recognition helix *(red cylinder)* are separated by exactly one turn of the DNA helix (3.4 nm). The other helix of the helix–turn–helix motif is colored *blue*, as in Figure 7–10. The lambda repressor and Cro proteins control bacteriophage lambda gene expression, and the tryptophan repressor and the catabolite activator protein (CAP) control the expression of sets of *E. coli* genes.

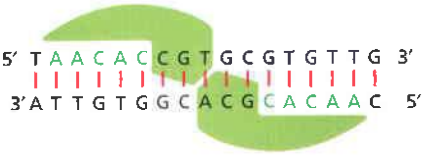


Figure 7–12 A specific DNA sequence recognized by the bacteriophage lambda Cro protein. The nucleotides labeled in *green* in this sequence are arranged symmetrically, allowing each half of the DNA site to be recognized in the same way by each protein monomer, also shown in *green*. See Figure 7–11 for the actual structure of the protein.

the bacterial gene regulatory proteins, providing one of the first indications that the principles of gene regulation established in bacteria are relevant to higher organisms as well. More than 60 homeodomain proteins have now been discovered in *Drosophila* alone, and homeodomain proteins have been identified in virtually all eucaryotic organisms that have been studied, from yeasts to plants to humans.

The structure of a homeodomain bound to its specific DNA sequence is shown in **Figure 7–13**. Whereas the helix–turn–helix motif of bacterial gene regulatory proteins is often embedded in different structural contexts, the helix–turn–helix motif of homeodomains is always surrounded by the same structure (which forms the rest of the homeodomain), suggesting that the motif is always presented to DNA in the same way. Indeed, structural studies have shown that a yeast homeodomain protein and a *Drosophila* homeodomain protein have very similar conformations and recognize DNA in almost exactly the same manner, although they are identical at only 17 of 60 amino acid positions (see Figure 3–13).

There Are Several Types of DNA-Binding Zinc Finger Motifs

The helix–turn–helix motif is composed solely of amino acids. A second important group of DNA-binding motifs includes one or more zinc atoms as structural components. Although all such zinc-coordinated DNA-binding motifs are called **zinc fingers**, this description refers only to their appearance in schematic drawings dating from their initial discovery (**Figure 7–14**A). Subsequent structural studies have shown that they fall into several distinct structural groups, two of which we consider here. The first type was initially discovered in the protein that activates the transcription of a eucaryotic ribosomal RNA gene. It has a simple structure, in which the zinc holds an α helix and a β sheet together (Figure 7–14B). This type of zinc finger is often found in tandem clusters so that the α helix of each can contact the major groove of the DNA, forming a nearly continuous stretch of α helices along the groove. In this way, a strong and specific DNA–protein interaction is built up through a repeating basic structural unit (**Figure 7–15**).

Another type of zinc finger is found in the large family of intracellular receptor proteins (discussed in detail in Chapter 15). It forms a different type of

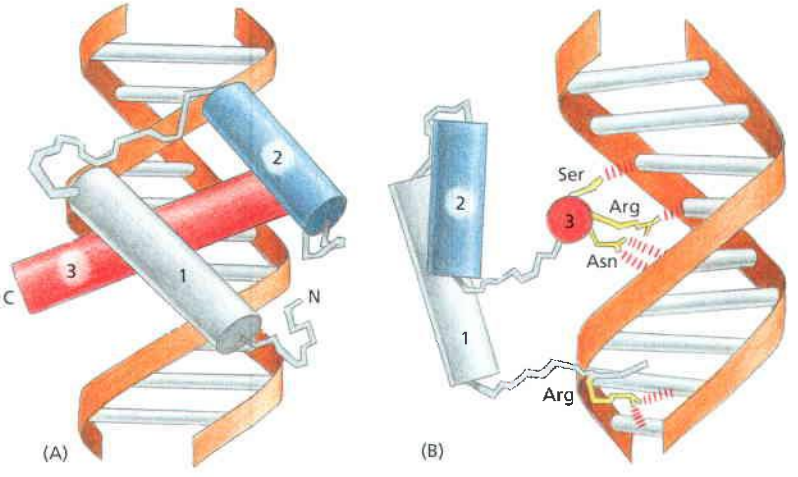


Figure 7–13 A homeodomain bound to its specific DNA sequence. Two different views of the same structure are shown. (A) The homeodomain is folded into three α helices, which are packed tightly together by hydrophobic interactions. The part containing helices 2 and 3 closely resembles the helix–turn–helix motif. (B) The recognition helix (helix 3, *red*) forms important contacts with the major groove of DNA. The asparagine (Asn) of helix 3, for example, contacts an adenine, as shown in Figure 7–9. A flexible arm attached to helix 1 forms contacts with nucleotide pairs in the minor groove. The homeodomain shown here is from a yeast gene regulatory protein, but it closely resembles homeodomains from many eucaryotic organisms. <ACGT> (Adapted from C. Wolberger et al., *Cell* 67:517–528, 1991. With permission from Elsevier.)

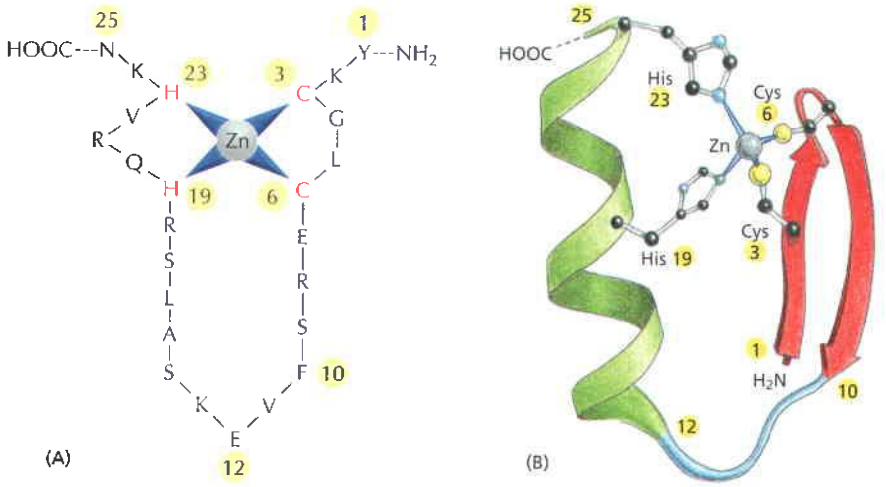


Figure 7–14 One type of zinc finger protein. This protein belongs to the Cys–Cys–His–His family of zinc finger proteins, named after the amino acids that grasp the zinc. (A) Schematic drawing of the amino acid sequence of a zinc finger from a frog protein of this class. (B) The three-dimensional structure of this same type of zinc finger is constructed from an antiparallel β sheet (amino acids 1 to 10) followed by an α helix (amino acids 12 to 24). The four amino acids that bind the zinc (Cys 3, Cys 6, His 19, and His 23) hold one end of the α helix firmly to one end of the β sheet. (Adapted from M.S. Lee et al., *Science* 245:635–637, 1989. With permission from AAAS.)

structure (similar in some respects to the helix–turn–helix motif) in which two α helices are packed together with zinc atoms (**Figure 7–16**). Like the helix–turn–helix proteins, these proteins usually form dimers that allow one of the two α helices of each subunit to interact with the major groove of the DNA. Although the two types of zinc finger structures discussed in this section are structurally distinct, they share two important features: both use zinc as a structural element, and both use an α helix to recognize the major groove of the DNA.

β sheets Can Also Recognize DNA

In the DNA-binding motifs discussed so far, α helices are the primary mechanism used to recognize specific DNA sequences. One large group of gene regulatory proteins, however, has evolved an entirely different recognition strategy. In this case, a two-stranded β sheet, with amino acid side chains extending from the sheet toward the DNA, reads the information on the surface of the major groove (**Figure 7–17**). As in the case of a recognition α helix, this β-sheet motif can be used to recognize many different DNA sequences; the exact DNA sequence recognized depends on the sequence of amino acids that make up the β sheet.

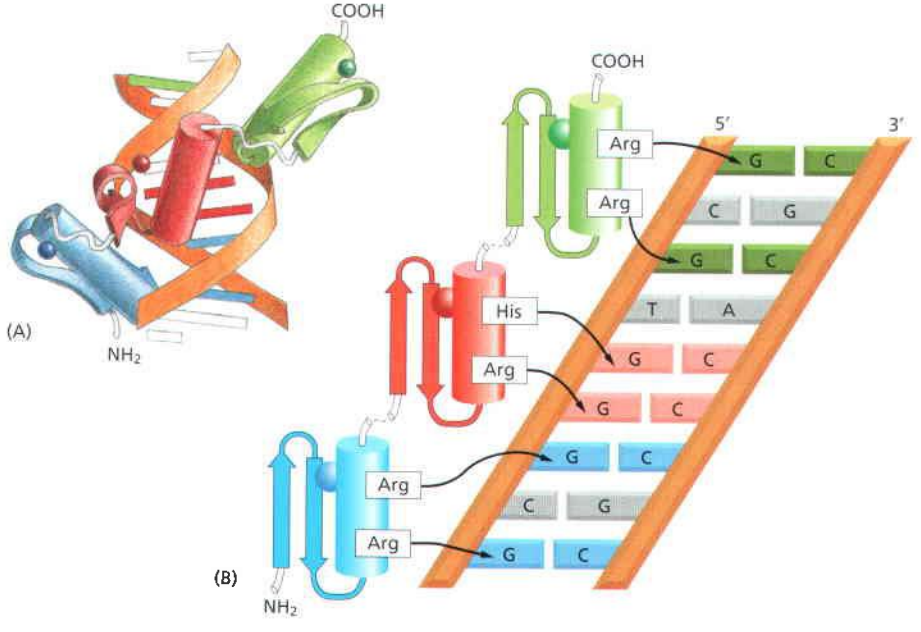


Figure 7–15 DNA binding by a zinc finger protein. (A) The structure of a fragment of a mouse gene regulatory protein bound to a specific DNA site. This protein recognizes DNA by using three zinc fingers of the Cys–Cys–His–His type (see Figure 7–14) arranged as direct repeats. <ATCT> (B) The three fingers have similar amino acid sequences and contact the DNA in similar ways. In both (A) and (B) the zinc atom in each finger is represented by a small sphere. (Adapted from N. Pavletich and C. Pabo, *Science* 252:810–817, 1991. With permission from AAAS.)

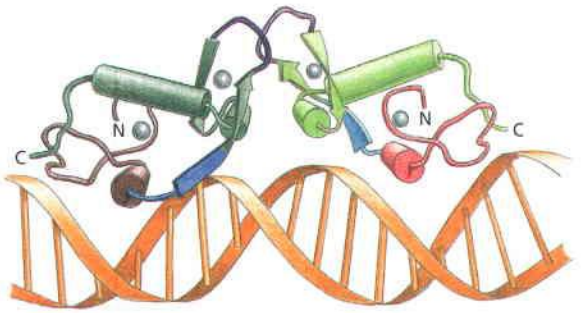


Figure 7–16 A dimer of the zinc finger domain of the intracellular receptor family bound to its specific DNA sequence. Each zinc finger domain contains two atoms of Zn (indicated by the small *gray spheres*); one stabilizes the DNA recognition helix (shown in *brown* in one subunit and *red* in the other), and one stabilizes a loop (shown in *purple*) involved in dimer formation. Each Zn atom is coordinated by four appropriately spaced cysteine residues. Like the helix–turn–helix proteins shown in Figure 7–11, the two recognition helices of the dimer are held apart by a distance corresponding to one turn of the DNA double helix. The specific example shown is a fragment of the glucocorticoid receptor. This is the protein through which cells detect and respond transcriptionally to the glucocorticoid hormones produced in the adrenal gland in response to stress. (Adapted from B.F. Luisi et al., *Nature* 352:497–505, 1991. With permission from Macmillan Publishers Ltd.)

Some Proteins Use Loops That Enter the Major and Minor Grooves to Recognize DNA

A few DNA-binding proteins use protruding peptide loops to read nucleotide sequences, rather than α helices and β sheets. For example, p53, a critical *tumor suppressor* in humans, recognizes nucleotide pairs from both the major and minor grooves using such loops (**Figure 7–18**). The normal role of the p53 protein is to tightly regulate cell growth and proliferation. Its importance can be appreciated by the fact that nearly half of all human cancers have acquired somatic mutations in the gene for p53; this step is key to the progression of many tumors, as we shall see in Chapter 20. Many of the p53 mutations observed in cancer cells destroy or alter its DNA-binding properties; indeed, Arg 248, which contacts the minor groove of DNA (see Figure 7–18) is the most frequently mutated p53 residue in human cancers.

The Leucine Zipper Motif Mediates Both DNA Binding and Protein Dimerization

Many gene regulatory proteins recognize DNA as homodimers, probably because, as we have seen, this is a simple way of achieving strong specific binding (see Figure 7–12). Usually, the portion of the protein responsible for dimerization is distinct from the portion that is responsible for DNA binding. One motif, however, combines these two functions elegantly and economically. It is called the **leucine zipper motif**, so named because of the way the two α helices, one from each monomer, are joined together to form a short coiled-coil (see Figure 3–9). The helices are held together by interactions between hydrophobic amino acid side chains (often on leucines) that extend from one side of each helix. Just beyond the dimerization interface the two α helices separate from each other to form a Y-shaped structure, which allows their side chains to contact the major groove of DNA. The dimer thus grips the double helix like a clothespin on a clothesline (**Figure 7–19**).

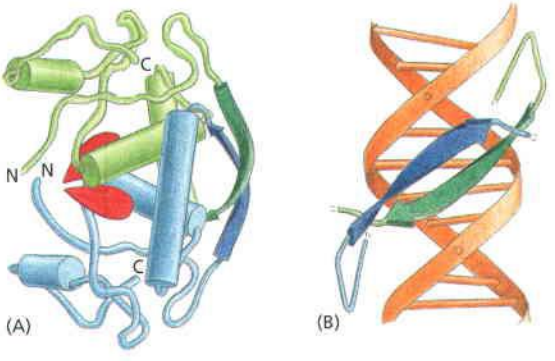


Figure 7–17 The bacterial Met repressor protein. The bacterial Met repressor regulates the genes encoding the enzymes that catalyze methionine synthesis. When this amino acid is abundant, it binds to the repressor, causing a change in the structure of the protein that enables it to bind to DNA tightly, shutting off the synthesis of the enzyme. (A) In order to bind to DNA tightly, the Met repressor must be complexed with *S*-adenosyl methionine, outlined in *red*. One subunit of the dimeric protein is shown in *green*, while the other is shown in *blue*. The two-stranded β sheet that binds to DNA is formed by one strand from each subunit and is shown in *dark green* and *dark blue*. (B) Simplified diagram of the Met repressor bound to DNA, showing how the two-stranded β sheet of the repressor binds to the major groove of DNA. For clarity, the other regions of the repressor have been omitted. (A, adapted from S. Phillips, *Curr. Opin. Struct. Biol.* 1:89–98, 1991, with permission from Elsevier; B, adapted from W. Somers and S. Phillips, *Nature* 359:387–393, 1992, with permission from Macmillan Publishers Ltd.)

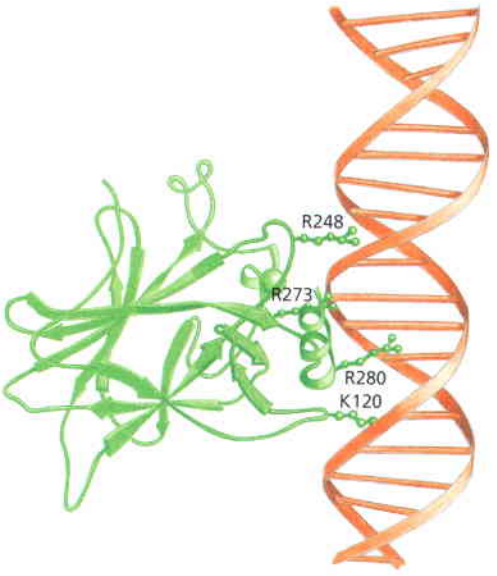


Figure 7–18 DNA recognition by the p53 protein. The most important DNA contacts are made by arginine 248 and lysine 120, which extend from the protruding loops entering the minor and major grooves. The folding of the p53 protein requires a zinc atom (shown as a sphere), but the way in which the zinc is grasped by the protein is completely different from that of the zinc finger proteins, described previously.

Heterodimerization Expands the Repertoire of DNA Sequences That Gene Regulatory Proteins Can Recognize

Many of the gene regulatory proteins we have seen thus far bind DNA as homodimers, that is, dimers made up of two identical subunits. However, many gene regulatory proteins can also associate with nonidentical partners to form heterodimers composed of two different subunits. Because heterodimers typically form from two proteins with distinct DNA-binding specificities, the mixing and matching of gene regulatory proteins in this way greatly expands the repertoire of DNA-binding specificities that these proteins can display. As illustrated in **Figure 7–20**, three distinct DNA-binding specificities could, in principle, be generated from two types of leucine zipper monomers, while six could be created from three types of monomers, and so on.

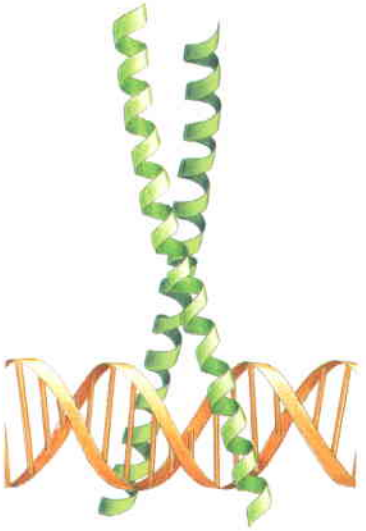


Figure 7–19 A leucine zipper dimer bound to DNA. Two α-helical DNA-binding domains *(bottom)* dimerize through their α-helical leucine zipper region *(top)* to form an inverted Y-shaped structure. Each arm of the Y is formed by a single α helix, one from each monomer, that mediates binding to a specific DNA sequence in the major groove of DNA. <TGTT> Each α helix binds to one-half of a symmetric DNA structure. The structure shown is of the yeast Gcn4 protein, which regulates transcription in response to the availability of amino acids in the environment. (Adapted from T.E. Ellenberger et al., *Cell* 71:1223–1237, 1992. With permission from Elsevier.)

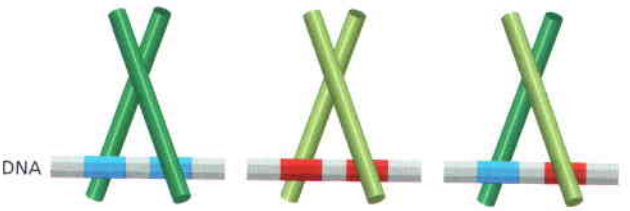


Figure 7–20 Heterodimerization of leucine zipper proteins can alter their DNA-binding specificity. Leucine zipper homodimers bind to symmetric DNA sequences, as shown in the left-hand and center drawings. These two proteins recognize different DNA sequences, as indicated by the *red* and *blue* regions in the DNA. The two different monomers can combine to form a heterodimer, which now recognizes a hybrid DNA sequence, composed from one *red* and one *blue* region.

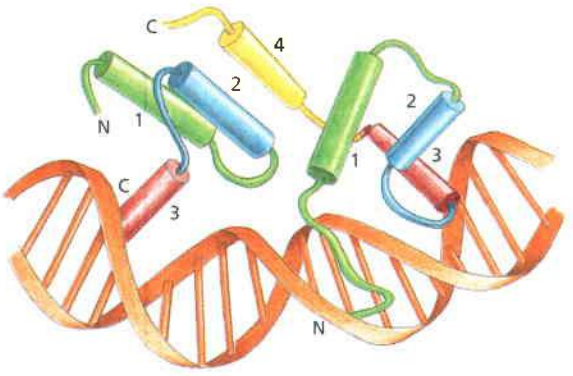


Figure 7–21 A heterodimer composed of two homeodomain proteins bound to its DNA recognition site. The *yellow* helix 4 of the protein on the right (Matα2) is unstructured in the absence of the protein on the left (Mata1), forming a helix only upon heterodimerization. The DNA sequence is recognized jointly by both proteins; some of the protein–DNA contacts made by Matα2 were shown in Figure 7–13. These two proteins are from budding yeast, where the heterodimer specifies a particular cell type (see Figure 7–65). The helices are numbered in accordance with Figure 7–13. (Adapted from T. Li et al., *Science* 270:262–269, 1995. With permission from AAAS.)

There are, however, limits to this promiscuity: for example, if all the many types of leucine zipper proteins in a typical eucaryotic cell formed heterodimers, the amount of "cross-talk" between the gene regulatory circuits of a cell would presumably be so great as to cause chaos. Whether or not a particular heterodimer can form depends on how well the hydrophobic surfaces of the two leucine zipper α helices mesh with each other, which in turn depends on the exact amino acid sequences of the two zipper regions. Thus, each leucine zipper protein in the cell can form dimers with only a small set of other leucine zipper proteins.

Heterodimerization is an example of **combinatorial control**, in which combinations of different proteins, rather than individual proteins, control a cell process. Heterodimerization as a mechanism for combinatorial control of gene expression occurs in many different types of gene regulatory proteins (**Figure 7–21**). Combinatorial control is a major theme that we shall encounter repeatedly in this chapter, and the formation of heterodimeric gene regulatory complexes is only one of many ways in which proteins work in combinations to control gene expression.

Certain combinations of gene regulatory proteins have become "hardwired" in the cell; for example, two distinct DNA-binding domains can, through gene rearrangements occurring over evolutionary time scales, become joined into a single polypeptide chain that displays a novel DNA-binding specificity (**Figure 7–22**).

The Helix–Loop–Helix Motif Also Mediates Dimerization and DNA Binding

Another important DNA-binding motif, related to the leucine zipper, is the **helix–loop–helix (HLH) motif**, which differs from the helix–turn–helix motif discussed earlier. An HLH motif consists of a short α helix connected by a loop to a second, longer α helix. The flexibility of the loop allows one helix to fold back

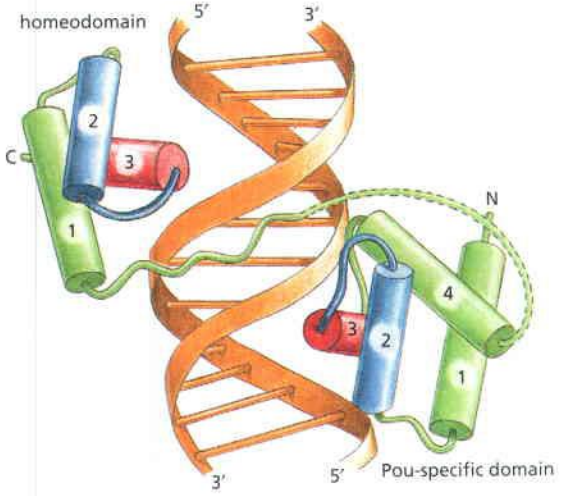


Figure 7–22 Two DNA-binding domains covalently joined by a flexible polypeptide. The structure shown (called a Pou-domain) consists of both a homeodomain and a helix–turn–helix structure joined by a flexible polypeptide "leash," indicated by the broken lines. A single gene encodes the entire protein, which is synthesized as a continuous polypeptide chain. The covalent joining of two structures in this way results in a large increase in the affinity of the protein for its specific DNA sequence compared with the DNA affinity of either separate structure. The group of mammalian gene regulatory proteins exemplified by this structure regulate the production of growth factors, immunoglobulins, and other molecules involved in development. The particular example shown is from the Oct1 protein. (Adapted from J.D. Klemm et al., *Cell* 77:21–32, 1994. With permission from Elsevier.)

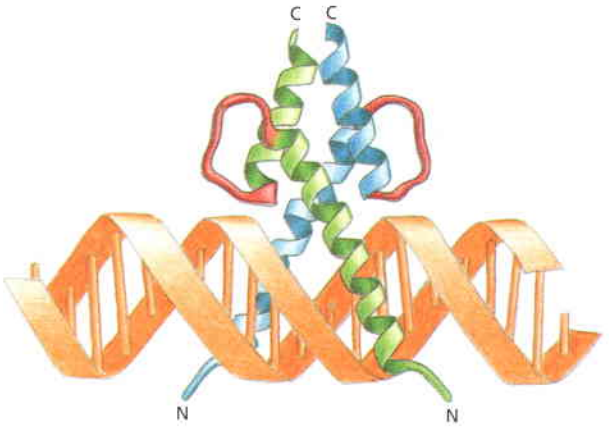


Figure 7–23 A helix–loop–helix (HLH) dimer bound to DNA. The two monomers are held together in a four-helix bundle: each monomer contributes two α helices connected by a flexible loop of protein *(red)*. A specific DNA sequence is bound by the two α helices that project from the four-helix bundle. (Adapted from A.R. Ferre-D'Amare et al., *Nature* 363:38–45, 1993. With permission from Macmillan Publishers Ltd.)

and pack against the other. As shown in **Figure 7–23**, this two-helix structure binds both to DNA and to the HLH motif of a second HLH protein. The second HLH protein can be the same (creating a homodimer) or different (creating a heterodimer). In either case, two α helices that extend from the dimerization interface make specific contacts with the DNA.

Several HLH proteins lack the α-helical extension responsible for binding to DNA. These truncated proteins can form heterodimers with full-length HLH proteins, but the heterodimers are unable to bind DNA tightly because they form only half of the necessary contacts. Thus, in addition to creating active dimers, heterodimerization provides cells with a widely used way to hold specific gene regulatory proteins in check (**Figure 7–24**).

It Is Not Yet Possible to Predict the DNA Sequences Recognized by All Gene Regulatory Proteins

The various DNA-binding motifs that we have discussed provide structural frameworks from which specific amino acid side chains extend to contact specific base pairs in the DNA. It is reasonable to ask, therefore, whether there is a simple amino acid–base pair recognition code: is a G–C base pair, for example, always contacted by a particular amino acid side chain? The answer is no, although certain types of amino acid-base interactions appear much more frequently than others (**Figure 7–25**). As we saw in Chapter 3, protein surfaces of virtually any shape and chemistry can be made from just 20 different amino acids, and a gene regulatory protein uses different patterns of these to create a surface that is precisely complementary to a particular DNA sequence. We know that the same base pair can thereby be recognized in many ways depending on its context (**Figure 7–26**). Nevertheless, molecular biologists are beginning to understand the principles of protein–DNA recognition well enough to design new proteins that will recognize a given DNA sequence.

Having outlined the general features of gene regulatory proteins, we turn to some of the methods that are now used to study them.

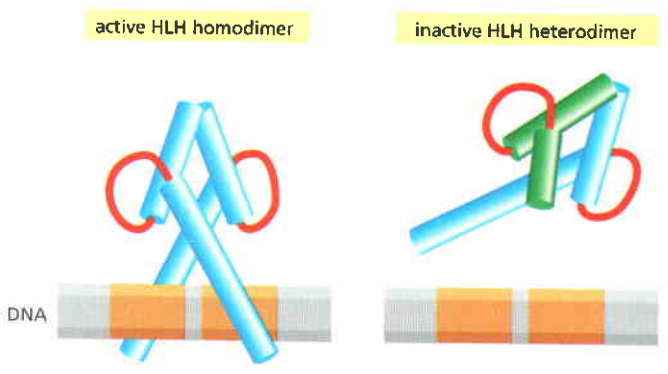


Figure 7–24 Inhibitory regulation by truncated HLH proteins. The HLH motif is responsible for both dimerization and DNA binding. On the *left*, an HLH homodimer recognizes a symmetric DNA sequence. On the *right*, the binding of a full-length HLH protein *(blue)* to a truncated HLH protein *(green)* that lacks the DNA-binding α helix generates a heterodimer that is unable to bind DNA tightly. If present in excess, the truncated protein molecule blocks the homodimerization of the full-length HLH protein and thereby prevents it from binding to DNA.

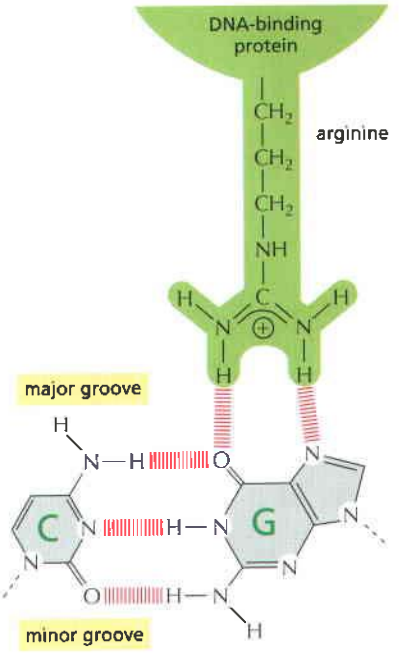


Figure 7–25 One of the most common protein–DNA interactions. Because of its specific geometry of hydrogen-bond acceptors (see Figure 7–7), the side chain of arginine unambiguously recognizes guanine. Figure 7–9 shows another common protein–DNA interaction.

A Gel-Mobility Shift Assay Readily Detects Sequence-Specific DNA-Binding Proteins

Genetic analyses, which provided a route to the gene regulatory proteins of bacteria, yeast, and *Drosophila*, are much more difficult in vertebrates. Therefore, the isolation of vertebrate gene regulatory proteins had to await the development of different approaches. Many of these approaches rely on the detection in a cell extract of a DNA-binding protein that specifically recognizes a DNA sequence known to control the expression of a particular gene. One of the most common ways to detect and study sequence-specific DNA-binding proteins is based on the effect of a bound protein on the migration of DNA molecules in an electric field.

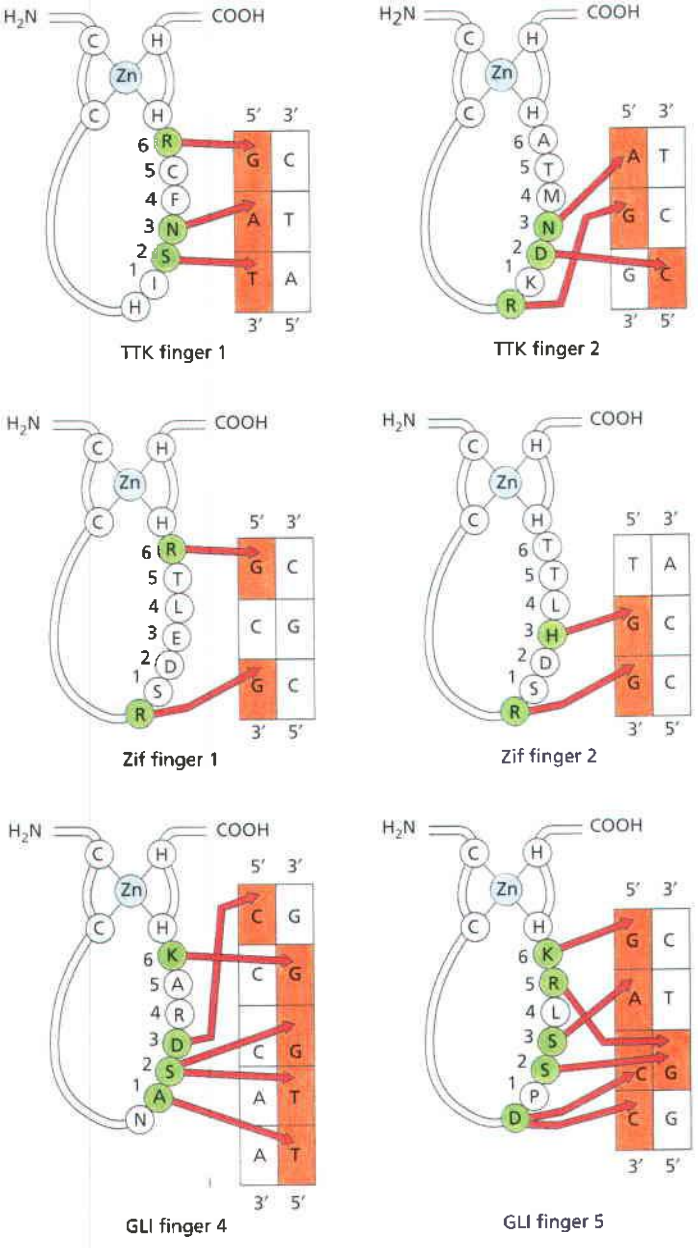


Figure 7–26 Summary of sequence-specific interactions between six different zinc fingers and their DNA recognition sequences. Even though all six Zn fingers have the same overall structure (see Figure 7–14), each binds to a different DNA sequence. The numbered amino acids form the α helix that recognizes DNA (Figures 7–14 and 7–15), and those that make sequence-specific DNA contacts are *green*. Bases contacted by protein are *orange*. Although arginine–guanine contacts are common (see Figure 7–25), guanine can also be recognized by serine, histidine, and lysine, as shown. Moreover, the same amino acid (serine, in this example) can recognize more than one base. Two of the Zn fingers depicted are from the TTK protein (a *Drosophila* protein that functions in development); two are from the mouse protein (Zif268) that was shown in Figure 7–15; and two are from a human protein (GL1) whose aberrant forms can cause certain types of cancers. (Adapted from C. Branden and J. Tooze, Introduction to Protein Structure, 2nd ed. New York: Garland Publishing, 1999.)

A DNA molecule is highly negatively charged and will therefore move rapidly toward a positive electrode when it is subjected to an electric field. When analyzed by polyacrylamide-gel electrophoresis (see p. 534), DNA molecules are separated according to their size because smaller molecules are able to penetrate the fine gel meshwork more easily than large ones. Protein molecules bound to a DNA molecule will cause it to move more slowly through the gel; in general, the larger the bound protein, the greater the retardation of the DNA molecule. This phenomenon provides the basis for the **gel-mobility shift assay**, which allows even trace amounts of a sequence-specific DNA-binding protein to be readily detected. In this assay, a short DNA fragment of specific length and sequence (produced either by DNA cloning or by chemical synthesis, as discussed in Chapter 8) is radioactively labeled and mixed with a cell extract; the mixture is then loaded onto a polyacrylamide gel and subjected to electrophoresis. If the DNA fragment corresponds to a chromosomal region where, for example, several sequence-specific proteins bind, autoradiography (see pp. 602–603) will reveal a series of DNA bands, each retarded to a different extent and representing a distinct DNA–protein complex. The proteins responsible for each band on the gel can then be separated from one another by subsequent fractionations of the cell extract (**Figure 7–27**). Once a sequence-specific DNA protein has been purified, the gel-mobility shift assay can be used to study the strength and specificity of its interactions with different DNA sequences, the lifetime of DNA–protein complexes, and other properties critical to the functioning of the protein in the cell.

DNA Affinity Chromatography Facilitates the Purification of Sequence-Specific DNA-Binding Proteins

A particularly powerful protein-purification method called **DNA affinity chromatography** can be used once the DNA sequence that a gene regulatory protein recognizes has been determined. A double-stranded oligonucleotide of the correct sequence is synthesized by chemical methods and linked to an insoluble porous matrix such as agarose; the matrix with the oligonucleotide attached is

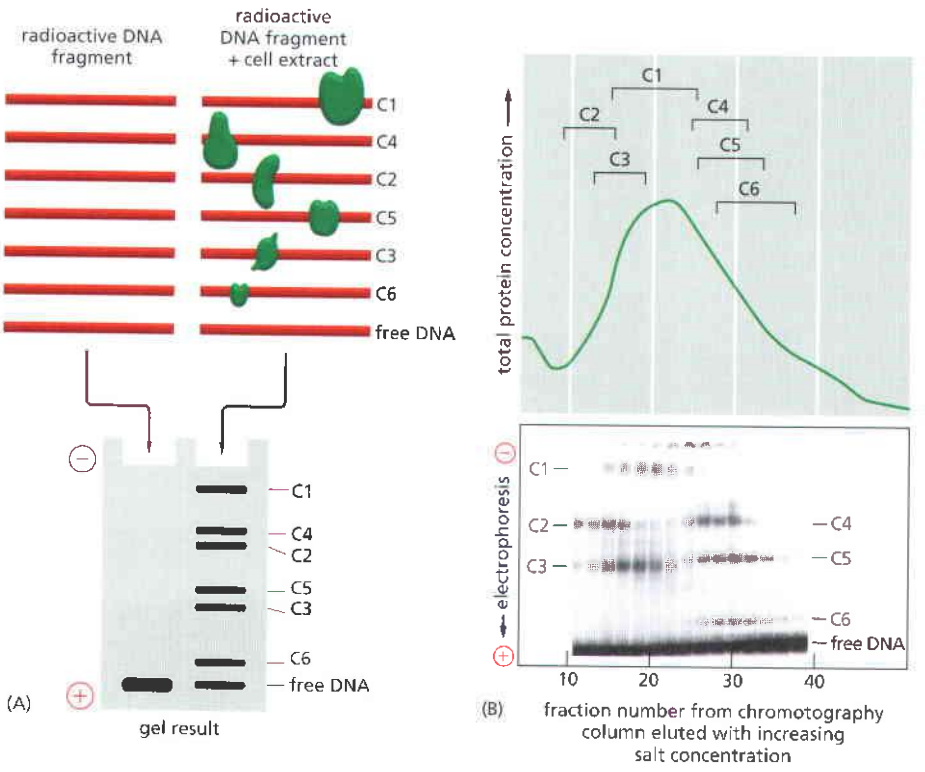


Figure 7–27 A gel-mobility shift assay. The principle of the assay is shown schematically in (A). In this example an extract of an antibody-producing cell line is mixed with a radioactive DNA fragment containing about 160 nucleotides of a regulatory DNA sequence from a gene encoding the light chain of the antibody made by the cell line. The effect of the proteins in the extract on the mobility of the DNA fragment is analyzed by polyacrylamide-gel electrophoresis followed by autoradiography. The free DNA fragments migrate rapidly to the bottom of the gel, while those fragments bound to proteins are retarded; the finding of six retarded bands suggests that the extract contains six different sequence-specific DNA-binding proteins (indicated as C1–C6) that bind to this DNA sequence. (For simplicity, any DNA fragments with more than one protein bound have been omitted from the figure.) In (B) a standard chromatographic technique (see pp. 512–513) was used to fractionate the extract *(top)*, and each fraction was mixed with the radioactive DNA fragment, applied to one lane of a polyacrylamide gel, and analyzed as in (A). (B, modified from C. Scheidereit, A. Heguy and R.G. Roeder, *Cell* 51:783–793, 1987. With permission from Elsevier.)

then used to construct a column that selectively binds proteins that recognize the particular DNA sequence (**Figure 7–28**). Purifications as great as 10,000-fold can be achieved by this means with relatively little effort.

Although most gene regulatory proteins are present at very low levels in the cell, enough pure protein can usually be isolated by affinity chromatography to obtain a partial amino acid sequence by mass spectrometry or other means (discussed in Chapter 8). If the complete genome sequence of the organism is known, the partial amino acid sequence can be used to identify the gene. The gene not only provides the complete amino acid sequence of the protein; it also provides the means to produce the protein in unlimited amounts through genetic engineering techniques, also discussed in Chapter 8.

The DNA Sequence Recognized by a Gene Regulatory Protein Can Be Determined Experimentally

Gene regulatory proteins can be discovered before the DNA sequence they recognize is known. For example, many of the *Drosophila* homeodomain proteins were discovered through the isolation of mutations that altered fly development. This allowed the genes encoding the proteins to be identified, and the proteins could then be overexpressed in cultured cells and easily purified. *DNA footprinting* is one method of determining the DNA sequences recognized by a gene regulatory protein once it has been purified. This strategy also requires a purified fragment of duplex DNA that contains somewhere within it a recognition site for the protein. Short recognition sequences can occur by chance on any long DNA fragment, although it is often necessary to use DNA corresponding to a regulatory region for a gene known to be controlled by the protein of interest. DNA footprinting is based on nucleases or chemicals that randomly cleave DNA at every phosphodiester bond. A bound gene regulatory protein blocks the phosphodiester bonds from attack, thereby revealing the protein's precise recognition site as a protected zone, or footprint (**Figure 7–29**).

A second way of determining the DNA sequences recognized by a gene regulatory protein requires no prior knowledge of what genes the protein might

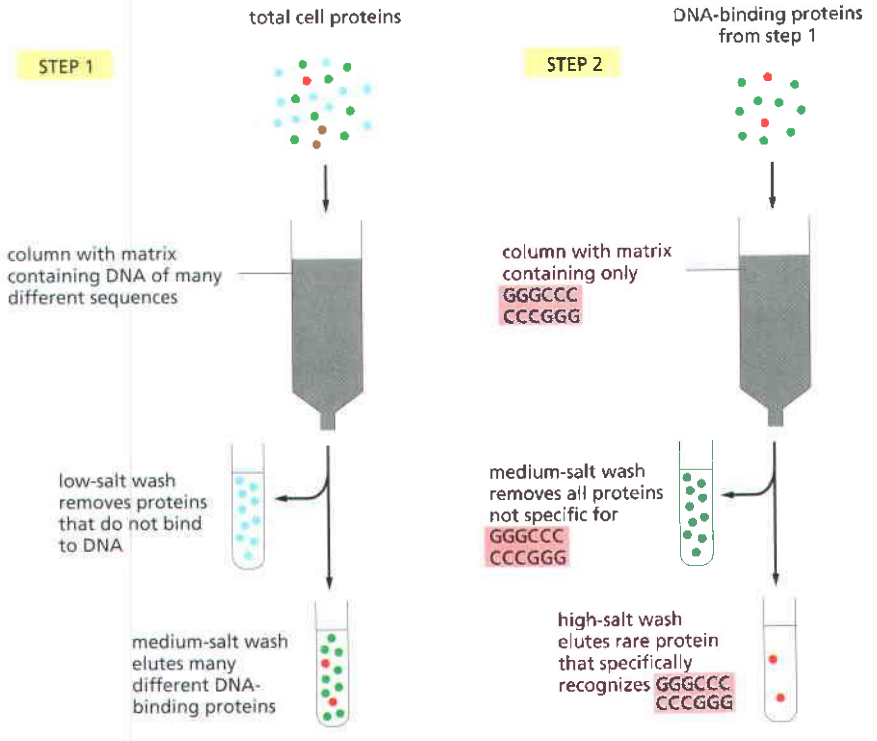


Figure 7–28 DNA affinity chromatography. In the first step, all the proteins that can bind DNA are separated from the remainder of the cell proteins on a column containing a huge number of different DNA sequences. Most sequence-specific DNA-binding proteins have a weak (nonspecific) affinity for bulk DNA and are therefore retained on the column. This affinity is due largely to ionic attractions, and the proteins can be washed off the DNA by a solution that contains a moderate concentration of salt. In the second step, the mixture of DNA-binding proteins is passed through a column that contains only DNA of a particular sequence. Typically, all the DNA-binding proteins will stick to the column, the great majority by nonspecific interactions. These are again eluted by solutions of moderate salt concentration, leaving on the column only those proteins (typically one or only a few) that bind specifically and therefore very tightly to the particular DNA sequence. These remaining proteins can be eluted from the column by solutions containing a very high concentration of salt.

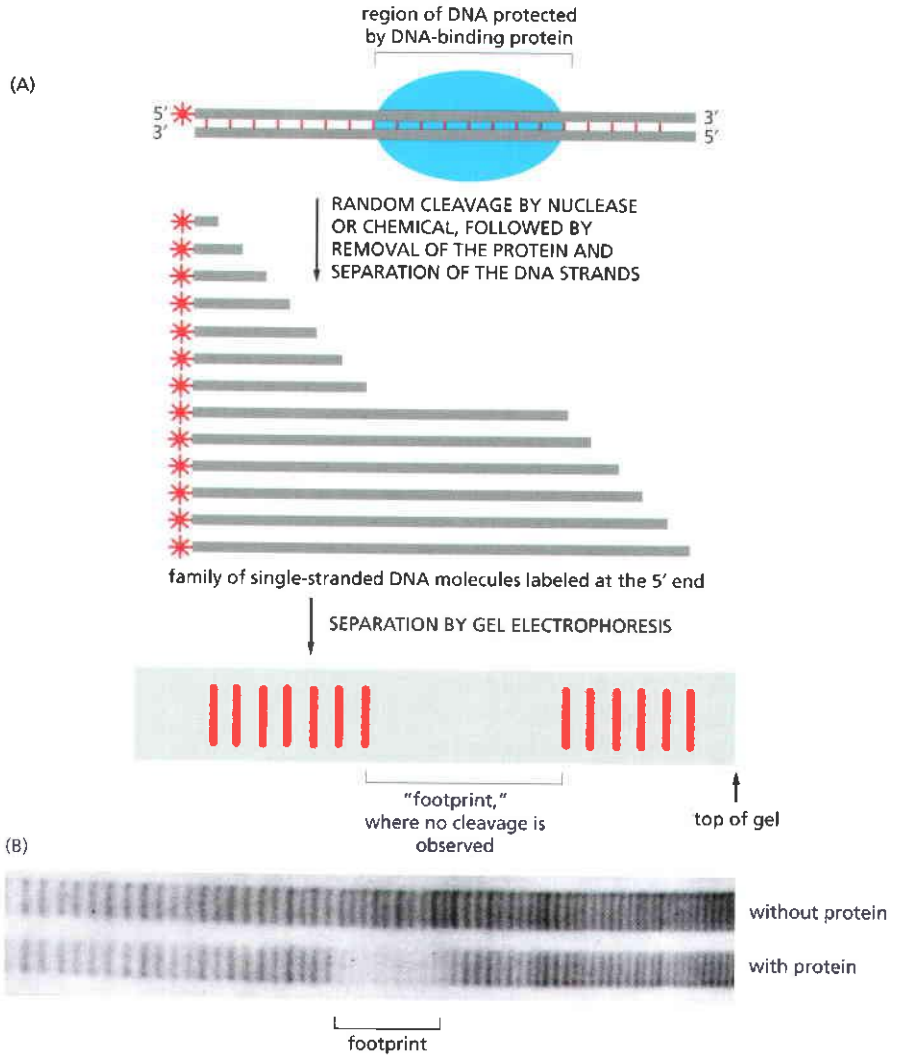


Figure 7–29 DNA footprinting. (A) Schematic of the method. A DNA fragment is labeled at one end with ^{32}P, a procedure described in Figure 8–34; next, the DNA is cleaved with a nuclease or chemical that makes random, single-stranded cuts. After the DNA molecule is denatured to separate its two strands, the resultant fragments from the labeled strand are separated on a gel and detected by autoradiography (see Figure 8–33). The pattern of bands from DNA cut in the presence of a DNA-binding protein is compared with that from DNA cut in its absence. When protein is present, it covers the nucleotides at its binding site and protects their phosphodiester bonds from cleavage. As a result, those labeled fragments that would otherwise terminate in the binding site are missing, leaving a gap in the gel pattern called a "footprint." In the example shown, the DNA-binding protein protects seven phosphodiester bonds from the DNA cleaving agent. (B) An actual footprint used to determine the binding site for a gene regulatory protein from humans. The cleaving agent was a small, iron-containing organic molecule that normally cuts at every phosphodiester bond with nearly equal frequency. (B, courtesy of Michele Sawadogo and Robert Roeder.)

regulate. Here, the purified protein is used to select, from a large, randomly generated pool of different short DNA fragments, only those that bind tightly to it. After several rounds of such selection, the nucleotide sequences of the tightly bound DNAs are determined, and a consensus DNA recognition sequence for the gene regulatory protein can be formulated (**Figure 7–30**). Once the DNA sequence recognized by a gene regulatory protein is known, computerized genome searches can identify candidate genes whose transcription the gene

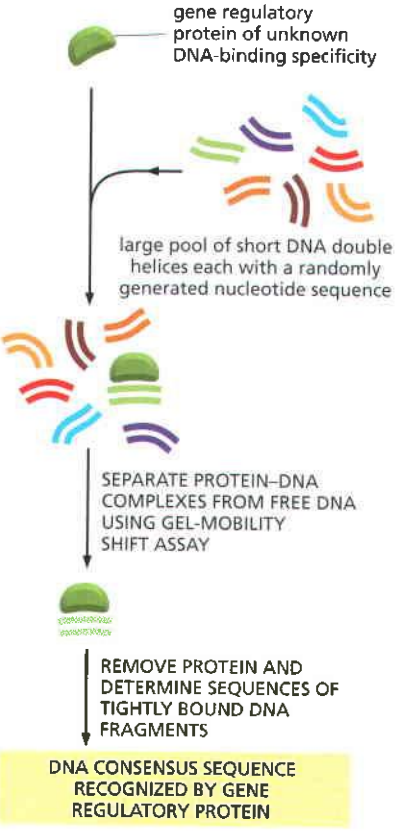


Figure 7–30 A method for determining the DNA sequence recognized by a gene regulatory protein. A purified gene regulatory protein is mixed with millions of different short DNA fragments, each with a different sequence of nucleotides. A collection of such DNA fragments can be produced by programming a DNA synthesizer, a machine that chemically synthesizes DNA of any desired sequence (discussed in Chapter 8). For example, there are 4^{11}, or approximately 4.2 million, possible sequences for a DNA fragment of 11 nucleotides. The double-stranded DNA fragments that bind tightly to the gene regulatory protein are then separated from the DNA fragments that fail to bind. One method for accomplishing this separation is through gel-mobility shifts, as illustrated in Figure 7–27. After separation of the DNA–protein complexes from the free DNA, the DNA fragments are removed from the protein and typically used for several additional rounds of the same selection process (not shown). The nucleotide sequences of those DNA fragments that remain through multiple rounds of binding and release can be determined, and a consensus DNA recognition sequence can thus be generated.



Figure 7–31 Phylogenetic footprinting. This example compares DNA sequences upstream of the same gene from five closely related yeasts; identical nucleotides are highlighted in *yellow*. Phylogenetic footprinting reveals DNA recognition sites for regulatory proteins, as they are typically more conserved than surrounding sequences. Only the region upstream of a particular gene is shown in this example, but the approach is typically used to analyze entire genomes. The gene regulatory proteins that bind to the site outlined in *red* are shown in Figure 7–21. Some of the shorter phylogenetic footprints in this example represent binding sites for additional gene regulatory proteins, not all of which have been identified. (From M. Kellis et al., *Nature* 423:241–254, 2003, with permission from Macmillan Publishers Ltd., and D.J. Galgoczy et al., *Proc. Natl Acad. Sci. U.S.A.* 101:18069–18074, 2004, with permission from National Academy of Sciences.)

regulatory protein of interest might control. However, this strategy is not foolproof. For example, many organisms produce a set of closely related gene regulatory proteins that recognize very similar DNA sequences, and this approach cannot resolve them. In most cases, predictions of the sites of action of gene regulatory proteins obtained from searching genome sequences must, in the end, be tested experimentally.

Phylogenetic Footprinting Identifies DNA Regulatory Sequences Through Comparative Genomics

The widespread availability of complete genome sequences provides a surprisingly simple method for identifying important regulatory sites on DNA, even when the gene regulatory protein that binds them is unknown. In this approach, genomes from several closely related species are compared. If the species are chosen properly, the protein-coding portions of the genomes will be very similar, but the regions between sequences that encode protein or RNA molecules will have diverged considerably, as most of this sequence is functionally irrelevant and therefore not constrained in evolution. Among the exceptions are the regulatory sequences that control gene transcription. These stand out as conserved islands in a sea of nonconserved nucleotides (**Figure 7–31**). Although the identity of the gene regulatory proteins that recognize the conserved DNA sequences must be determined by other means, phylogenetic footprinting is a powerful method for identifying many of the DNA sequences that control gene expression.

Chromatin Immunoprecipitation Identifies Many of the Sites That Gene Regulatory Proteins Occupy in Living Cells

A gene regulatory protein will not occupy all of its potential DNA-binding sites in the genome at a particular time. Under some conditions, the protein may not be synthesized, and so will be absent from the cell; it may be present but lacking a heterodimer partner; or it may be excluded from the nucleus until an appropriate signal is received from the cell's environment. Even if the gene regulatory

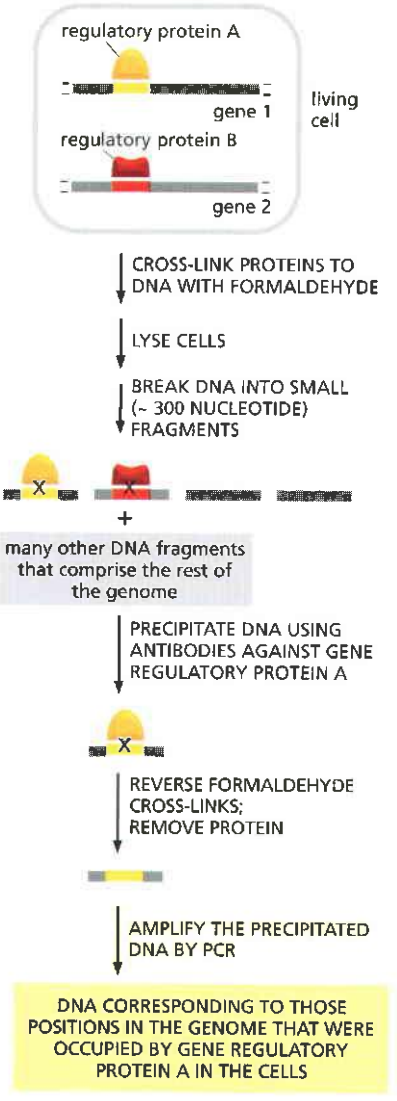


Figure 7–32 Chromatin immunoprecipitation. This method allows the identification of all the sites in a genome that a gene regulatory protein occupies *in vivo*. For the amplification of DNA by a polymerase chain reaction (PCR), see Figure 8–45. The identities of the precipitated, amplified DNA fragments can be determined by hybridizing the mixture of fragments to DNA microarrays, as described in Chapter 8.

protein is present in the nucleus and is competent to bind DNA, components of chromatin or other gene regulatory proteins that can bind to the same or overlapping DNA sequences may occlude many of its potential binding sites on DNA.

Chromatin immunoprecipitation provides one way of empirically determining the sites on DNA that a given gene regulatory protein occupies under a particular set of conditions (**Figure 7–32**). In this approach, proteins are covalently cross-linked to DNA in living cells, the cells are broken open, and the DNA is mechanically sheared into small fragments. Antibodies directed against a given gene regulatory protein are then used to purify DNA that became covalently cross-linked to that protein in the cell. If this DNA is hybridized to microarrays that contain the entire genome displayed as a series of discrete DNA fragments (see Figure 8–73), the precise genomic location of each precipitated DNA fragment can be determined. In this way, all the sites occupied by the gene regulatory protein in the original cells can be mapped on the cell's genome (**Figure 7–33**).

Chromatin immunoprecipitation is also routinely used to identify the positions along a genome that are packaged by the various types of modified histones (discussed in Chapter 4). In this case, antibodies specific to the particular histone modification of interest are employed.

Summary

Gene regulatory proteins recognize short stretches of double-helical DNA of defined sequence and thereby determine which of the thousands of genes in a cell will be transcribed. Thousands of gene regulatory proteins have been identified in a wide variety of organisms. Although each of these proteins has unique features, most bind to DNA as homodimers or heterodimers and recognize DNA through one of a small number of structural motifs. The common motifs include the helix–turn–helix, the homeodomain, the leucine zipper, the helix–loop–helix, and zinc fingers of several types. The precise amino acid sequence that is folded into a motif determines the particular DNA sequence that a gene regulatory protein recognizes. Heterodimerization increases the range of DNA sequences that can be recognized. Powerful techniques are now available for identifying and isolating these proteins, the genes that encode them, and the DNA sequences they recognize, and for mapping all of the genes that they regulate on a genome.

HOW GENETIC SWITCHES WORK

In the previous section, we described the basic components of genetic switches: gene regulatory proteins and the specific DNA sequences that these proteins recognize. We shall now discuss how these components operate to turn genes on and off in response to a variety of signals.

In the mid-twentieth century, the idea that genes could be switched on and off was revolutionary. This concept was a major advance, and it came originally from the study of how *E. coli* bacteria adapt to changes in the composition of their growth medium. Parallel studies of the lambda bacteriophage led to many of the same conclusions and helped to establish the underlying mechanism. Many of the same principles apply to eucaryotic cells. However, the enormous complexity of gene regulation in higher organisms, combined with the packaging

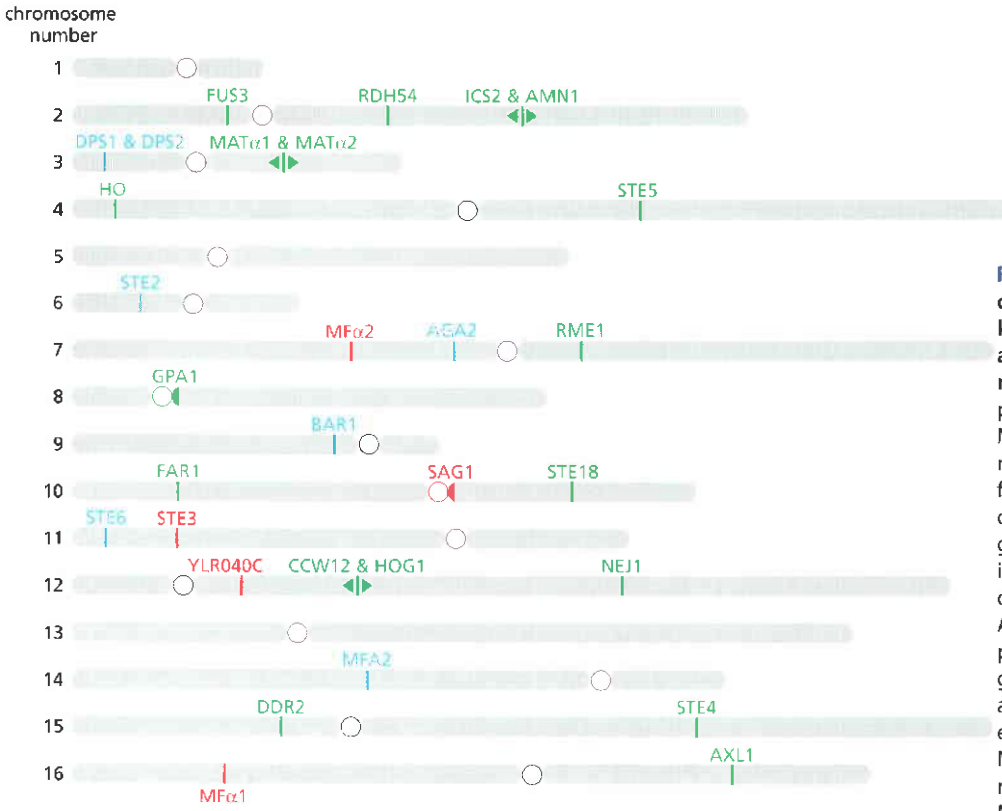


Figure 7–33 A gene regulatory circuit: the complete set of genes controlled by three key regulatory proteins in budding yeast, as deduced from the DNA sites where the regulatory proteins bind. The regulatory proteins—called Mata1, Matα1, and Matα2—specify the two different haploid mating types (analogous to male and female gamates) of this unicellular organism. The 16 chromosomes in the yeast genome are shown *(gray)*, with colored bars indicating sites where various combinations of the three regulatory proteins bind. Above each binding site is the name of the protein product of the regulated target gene. Matα1, acting in a complex with another protein, Mcm1, activates expression of the genes marked in *red;* Matα2, acting in a complex with Mcm1, represses the genes marked in *blue;* and Mata1 in a complex with Matα2 represses the genes marked in *green* (see Figures 7–21 and 7–65). *Double arrowheads* represent divergently transcribed genes, which are controlled by the indicated gene regulatory proteins. This complete map of bound regulatory proteins was determined using a combination of genome-wide chromatin immunoprecipitation (see Figure 7–32) and phylogenetic footprinting (see Figure 7–29). Such determinations of complete transcriptional circuits show that transcriptional networks are not infinitely complex, although they may appear that way initially. This type of study also helps to reveal the overall logic of the transcriptional circuits used by modern cells. (From D.J. Galgoczy et al., *Proc. Natl Acad. Sci. U.S.A.* 101:18069–18074, 2004. With permission from National Academy of Sciences.)

of their DNA into chromatin, creates special challenges and some novel opportunities for control—as we shall see. We begin with the simplest example—an on–off switch in bacteria that responds to a single signal.

The Tryptophan Repressor Is a Simple Switch That Turns Genes On and Off in Bacteria

The chromosome of the bacterium *E. coli*, a single-celled organism, is a single circular DNA molecule of about 4.6×10^6 nucleotide pairs. This DNA encodes approximately 4300 proteins, although the cell makes only a fraction of these at any one time. The expression of many genes is regulated according to the available food in the environment. This is illustrated by the five *E. coli* genes that code for enzymes that manufacture the amino acid tryptophan. These genes are arranged as a single **operon**; that is, they are adjacent to one another on the chromosome and are transcribed from a single *promoter* as one long mRNA molecule (**Figure 7–34**). But when tryptophan is present in the growth medium and enters the cell (when the bacterium is in the gut of a mammal that has just eaten a meal of protein, for example), the cell no longer needs these enzymes and shuts off their production.

The molecular basis for this switch is understood in considerable detail. As described in Chapter 6, a promoter is a specific DNA sequence that directs RNA polymerase to bind to DNA, to open the DNA double helix, and to begin

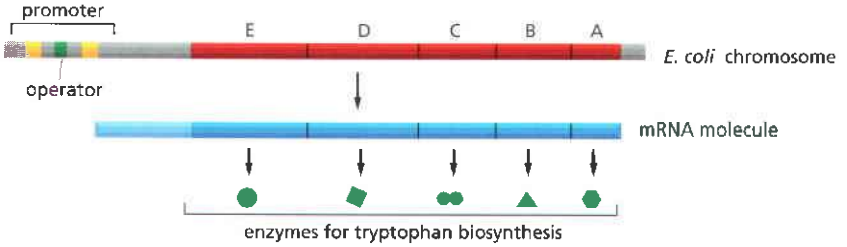
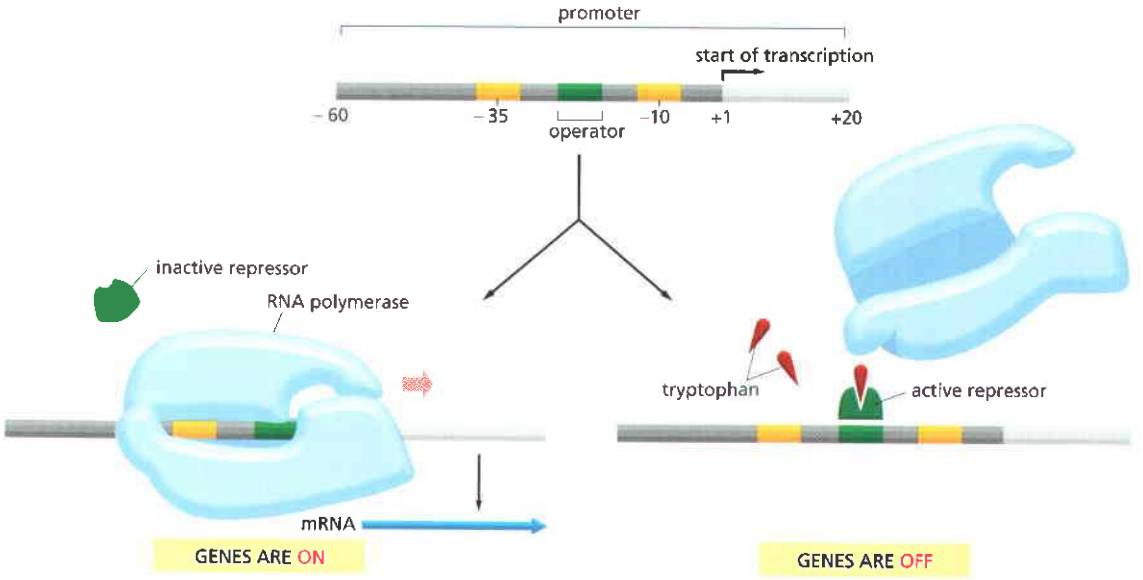


Figure 7–34 The clustered genes in *E. coli* that code for enzymes that manufacture the amino acid tryptophan. These five genes of the *Trp* operon—denoted as *TrpA, B, C, D,* and *E*—are transcribed as a single mRNA molecule, which allows their expression to be controlled coordinately. Clusters of genes transcribed as a single mRNA molecule are common in bacteria. Each such cluster is called an operon.



synthesizing an RNA molecule. Within the promoter that directs transcription of the tryptophan biosynthetic genes lies a regulator element called an **operator** (see Figure 7–34). This is simply a short region of regulatory DNA of defined nucleotide sequence that is recognized by a repressor protein, in this case the **tryptophan repressor**, a member of the helix–turn–helix family (see Figure 7–11). The promoter and operator are arranged so that when the tryptophan repressor occupies the operator, it blocks access to the promoter by RNA polymerase, thereby preventing expression of the tryptophan-producing enzymes (**Figure 7–35**).

The block to gene expression is regulated in an ingenious way: to bind to its operator DNA, the repressor protein has to have two molecules of the amino acid tryptophan bound to it. As shown in **Figure 7–36**, tryptophan binding tilts the helix–turn–helix motif of the repressor so that it is presented properly to the DNA major groove; without tryptophan, the motif swings inward and the protein is unable to bind to the operator. Thus, the tryptophan repressor and operator form a simple device that switches production of the tryptophan biosynthetic enzymes on and off according to the availability of free tryptophan.

Figure 7–35 Switching the tryptophan genes on and off. If the level of tryptophan inside the cell is low, RNA polymerase binds to the promoter and transcribes the five genes of the tryptophan (*Trp*) operon. If the level of tryptophan is high, however, the tryptophan repressor is activated to bind to the operator, where it blocks the binding of RNA polymerase to the promoter. Whenever the level of intracellular tryptophan drops, the repressor releases its tryptophan and becomes inactive, allowing the polymerase to begin transcribing these genes. The promoter includes two key blocks of DNA sequence information, the −35 and −10 regions highlighted in *yellow* (see Figure 6–12).

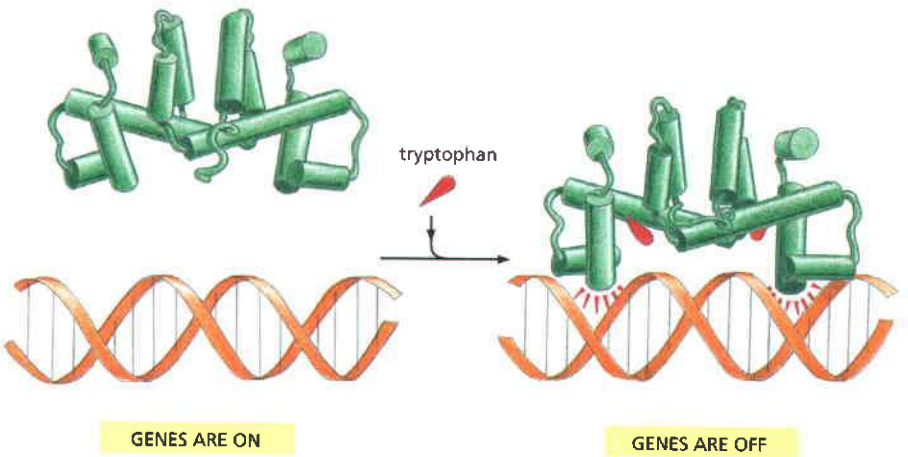


Figure 7–36 The binding of tryptophan to the tryptophan repressor protein changes its conformation. This structural change enables this gene regulatory protein to bind tightly to a specific DNA sequence (the operator), thereby blocking transcription of the genes encoding the enzymes required to produce tryptophan (the *Trp* operon). The three-dimensional structure of this bacterial helix–turn–helix protein, as determined by x-ray diffraction with and without tryptophan bound, is illustrated. Tryptophan binding increases the distance between the two recognition helices in the homodimer, allowing the repressor to fit snugly on the operator. (Adapted from R. Zhang et al., *Nature* 327:591–597, 1987. With permission from Macmillan Publishers Ltd.)

Because the active, DNA-binding form of the protein serves to turn genes off, this mode of gene regulation is called **negative control**, and the gene regulatory proteins that function in this way are called *transcriptional repressors* or *gene repressor proteins.*

Transcriptional Activators Turn Genes On

We saw in Chapter 6 that purified *E. coli* RNA polymerase (including its σ subunit) can bind to a promoter and initiate DNA transcription. Many bacterial promoters, however, are only marginally functional on their own, either because they are recognized poorly by RNA polymerase or because the polymerase has difficulty opening the DNA helix and beginning transcription. In either case these poorly functioning promoters can be rescued by gene regulatory proteins that bind to a nearby site on the DNA and contact the RNA polymerase in a way that dramatically increases the probability that a transcript will be initiated. Because the active, DNA-binding form of such a protein turns genes on, this mode of gene regulation is called **positive control**, and the gene regulatory proteins that function in this manner are known as *transcriptional activators* or *gene activator proteins.* In some cases, bacterial gene activator proteins aid RNA polymerase in binding to the promoter by providing an additional contact surface for the polymerase. In other cases, they contact RNA polymerase and facilitate its transition from the initial DNA-bound conformation of polymerase to the actively transcribing form by stabilizing a transition state of the enzyme. Like repressors, gene activator proteins must be bound to DNA to exert their effects. In this way, each regulatory protein acts selectively, controlling only those genes that bear a DNA sequence recognized by it.

DNA-bound activator proteins can increase the rate of transcription initiation up to 1000-fold, a value consistent with a relatively weak and nonspecific interaction between the activator and RNA polymerase. For example, a 1000-fold change in the affinity of RNA polymerase for its promoter corresponds to a change in ΔG of ~4 kcal/mole, which could be accounted for by just a few weak, noncovalent bonds. Thus gene activator proteins can work simply by providing a few favorable interactions that help to attract RNA polymerase to the promoter.

As in negative control by a transcriptional repressor, a transcriptional activator can operate as part of a simple on–off genetic switch. The bacterial activator protein *CAP (catabolite activator protein),* for example, activates genes that enable *E. coli* to use alternative carbon sources when glucose, its preferred carbon source, is unavailable. Falling levels of glucose cause an increase in the intracellular signaling molecule cyclic AMP, which binds to the CAP protein, enabling it to bind to its specific DNA sequence near target promoters and thereby turn on the appropriate genes. In this way the expression of a target gene is switched on or off, depending on whether cyclic AMP levels in the cell are high or low, respectively. **Figure 7–37** summarizes the different ways that positive and negative control can be used to regulate genes.

Transcriptional activators and transcriptional repressors are similar in design. The tryptophan repressor and the transcriptional activator CAP, for example, both use a helix–turn–helix motif (see Figure 7–11) and both require a small cofactor in order to bind DNA. In fact, some bacterial proteins (including CAP and the bacteriophage lambda repressor) can act as either activators or repressors, depending on the exact placement of the DNA sequence they recognize in relation to the promoter: if the binding site for the protein overlaps the promoter, the polymerase cannot bind and the protein acts as a repressor (**Figure 7–38**).

A Transcriptional Activator and a Transcriptional Repressor Control the *Lac* Operon

More complicated types of genetic switches combine positive and negative controls. The *Lac operon* in *E. coli,* for example, unlike the *Trp operon,* is under both

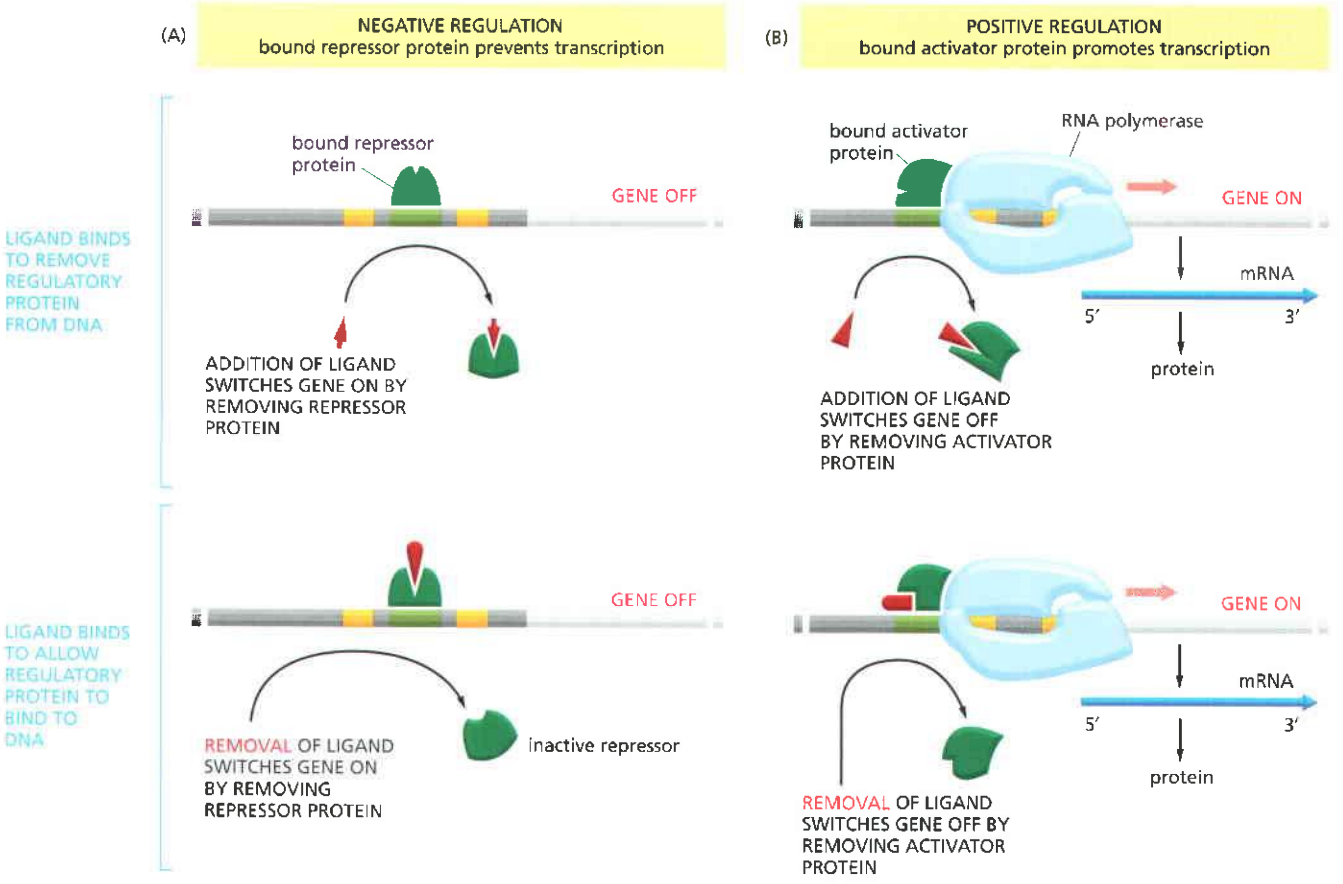
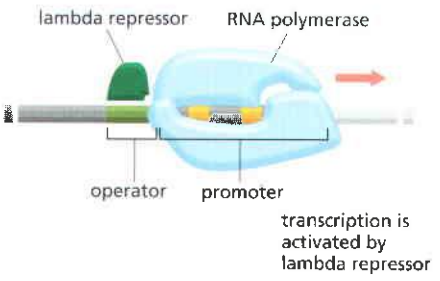


Figure 7–37 Summary of the mechanisms by which specific gene regulatory proteins control gene transcription in procaryotes. (A) Negative regulation; (B) positive regulation. Note that the addition of an "inducing" ligand can turn on a gene either by removing a gene repressor protein from the DNA *(upper left panel)* or by causing a gene activator protein to bind *(lower right panel)*. Likewise, the addition of an "inhibitory" ligand can turn off a gene either by removing a gene activator protein from the DNA *(upper right panel)* or by causing a gene repressor protein to bind *(lower left panel)*.

negative and positive transcriptional controls by the Lac repressor protein and CAP, respectively. The *Lac* operon codes for proteins required to transport the disaccharide lactose into the cell and to break it down. CAP, as we have seen, enables bacteria to use alternative carbon sources such as lactose in the absence of glucose. It would be wasteful, however, for CAP to induce expression of the *Lac* operon if lactose is not present, and the Lac repressor ensures that the *Lac* operon is shut off in the absence of lactose. This arrangement enables the control region of the *Lac* operon to respond to and integrate two different signals, so that the operon is highly expressed only when two conditions are met: lactose must be present and glucose must be absent. In any of the other three possible signal combinations, the cluster of genes is held in the off state (**Figure 7–39**).

The simple logic of this genetic switch first attracted the attention of biologists over 50 years ago. As explained above, the molecular basis of the switch was uncovered by a combination of genetics and biochemistry, providing the first glimpse into how gene expression is controlled.



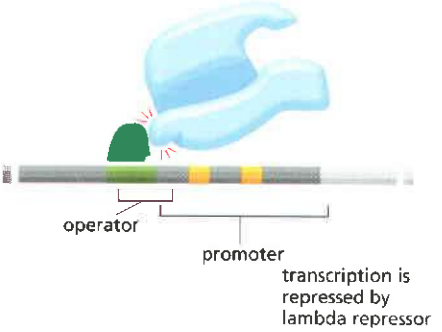


Figure 7–38 Some bacterial gene regulatory proteins can act as either a transcriptional activator or a repressor, depending on the precise placement of their DNA-binding sites. An example is the bacteriophage lambda repressor. For some genes, the protein acts as a transcriptional activator by providing a favorable contact for RNA polymerase *(top)*. At other genes *(bottom)*, the operator is located one base pair closer to the promoter, and, instead of helping polymerase, the repressor now competes with it for binding to the DNA. The lambda repressor recognizes its operator by a helix–turn–helix motif, as shown in Figure 7–11.

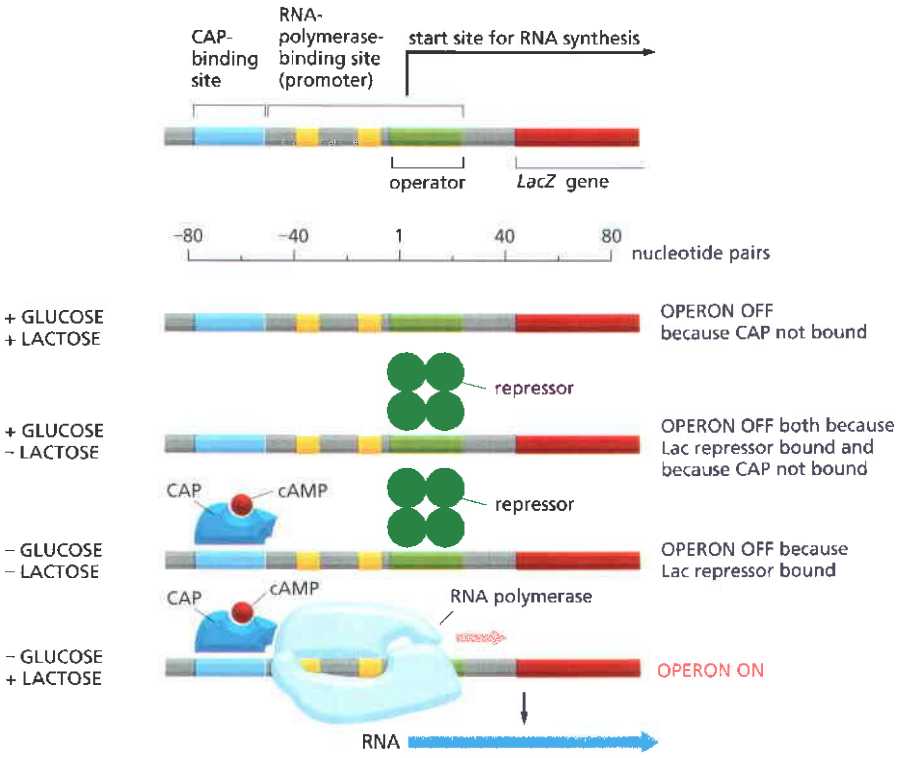


Figure 7–39 Dual control of the *Lac* operon. Glucose and lactose levels control the initiation of transcription of the *Lac* operon through their effects on CAP and the *Lac* repressor protein, respectively. *LacZ*, the first gene of the *Lac* operon, encodes the enzyme β-galactosidase, which breaks down the disaccharide lactose to galactose and glucose. Lactose addition increases the concentration of allolactose, an isomer of lactose, which binds to the repressor protein and removes it from the DNA. Glucose addition decreases the concentration of cyclic AMP; because cyclic AMP no longer binds to CAP, this gene activator protein dissociates from the DNA, turning off the operon.

This figure summarizes the essential features of the *Lac* operon, but in reality the situation is more complex. There are several Lac repressor binding sites located at different positions along the DNA. Although the one illustrated exerts the greatest effect, the others are required for full repression (see Figure 7–40). In addition, expression of the *Lac* operon never completely shuts down. A small amount of the enzyme β-galactosidase is required to convert lactose to allolactose, thereby permitting the Lac repressor to be inactivated when lactose is added to the growth medium.

DNA Looping Occurs During Bacterial Gene Regulation

The control of the Lac operon as shown in Figure 7–39 is simple and economical, but the continued study of this and other examples of bacterial gene regulation revealed a new feature of gene regulation, known as *DNA looping*. The *Lac* operon was originally thought to contain a single operator, but subsequent work revealed additional, secondary operators located nearby. A single tetrameric molecule of the Lac repressor can bind two operators simultaneously, looping out the intervening DNA. The ability to bind simultaneously to two operators strengthens the overall interaction of the Lac repressor with DNA and thereby leads to greater levels of repression in the cell (**Figure 7–40**).

DNA looping also allows two different proteins bound along a DNA double helix to contact one another readily. The DNA can be thought of as a tether, helping one DNA-bound protein interact with another even though thousands of nucleotide pairs may separate the binding sites for the two proteins (**Figure 7–41**). We shall see below that DNA looping is especially important in eucaryotic

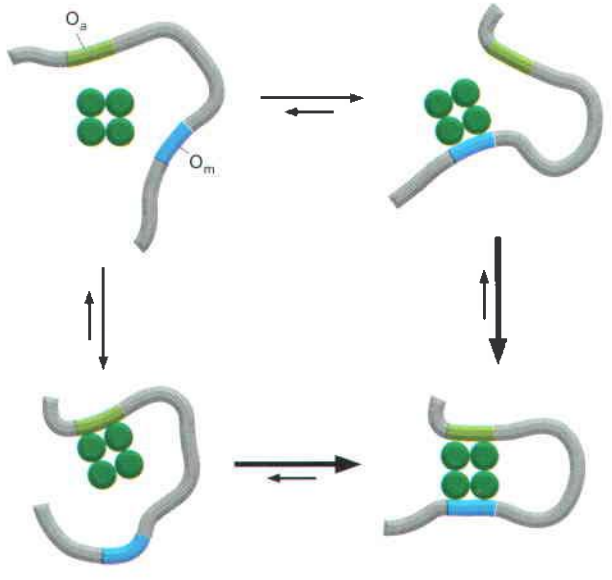


Figure 7–40 DNA looping can stabilize protein–DNA interactions. The Lac repressor, a tetramer, can simultaneously bind to two operators. The *Lac* operon has a total of three operators, but for simplicity, only two are shown here, the main operator (O_m) and an auxiliary operator (O_a). The figure shows all the possible states of the Lac repressor bound to these two operators. At the concentrations of Lac repressor in the cell, and in the absence of lactose, the state in the lower right is the most stable, and to dissociate completely from the DNA, the Lac repressor must first pass through an intermediate where it is bound to only a single operator. In these states, the local concentration of the repressor is very high in relation to the free operator, and the reaction to the double-bound form is favored over the dissociation reaction. In this way, even a low-affinity site (O_a) can increase the occupancy of a high-affinity site (O_m) and give higher levels of gene repression in the cell. (Adapted from J.M.G. Vilar and S. Leibler, *J. Mol. Biol.* 331:981–989, 2003. With permission from Academic Press.)

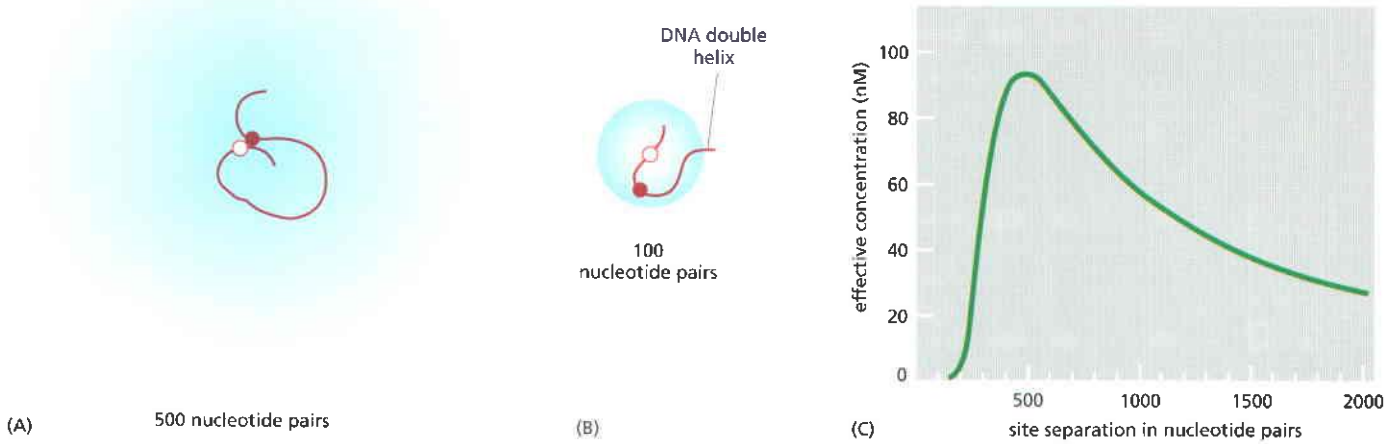


Figure 7–41 Binding of two proteins to separate sites on the DNA double helix can greatly increase their probability of interacting. (A) The tethering of one protein to the other via an intervening DNA loop of 500 nucleotide pairs increases their frequency of collision. The intensity of the *blue* coloring at each point in space indicates the probability that the *red* protein will be located at that distance from the *white* protein. (B) The flexibility of DNA is such that an average sequence makes a smoothly graded 90° bend (a curved turn) about once every 200 nucleotide pairs. Thus, when only 100 nucleotide pairs tethers two proteins, the contact between those proteins is relatively restricted. In such cases the protein interaction is facilitated when the two protein-binding sites are separated by a multiple of about 10 nucleotide pairs, which places both proteins on the same side of the DNA helix (which has about 10 nucleotides per turn) and thus on the inside of the DNA loop, where they can best reach each other. (C) The theoretical effective concentration of the *red* protein at the site where the *white* protein is bound, as a function of their separation. Experiments suggest that the actual effective concentrations at short distances are greater than those predicted here. (C, courtesy of Gregory Bellomy, modified from M.C. Mossing and M.T. Record, *Science* 233:889–892, 1986. With permission from AAAS.)

gene regulation. However, it also plays crucial roles in many examples of bacterial gene regulation in addition to that of the *Lac* operon. For example, DNA looping readily allows the bacterial gene activator protein NtrC to contact RNA polymerase directly even though the two proteins are bound several hundred nucleotide pairs apart (Figure 7–42).

Bacteria Use Interchangeable RNA Polymerase Subunits to Help Regulate Gene Transcription

We have seen the importance of gene regulatory proteins that bind to sequences of DNA and signal to RNA polymerase whether or not to start the synthesis of an

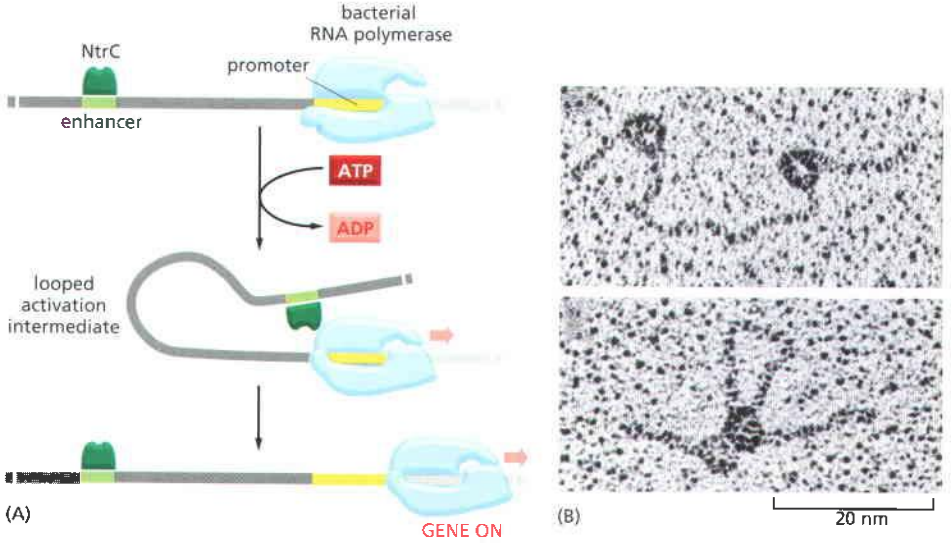


Figure 7–42 Gene activation at a distance. (A) NtrC is a bacterial gene regulatory protein that activates transcription by directly contacting RNA polymerase and causing a transition between the initial DNA-bound form of the polymerase and the transcriptionally competent form (discussed in Chapter 6). As indicated, the transition stimulated by NtrC requires the energy from ATP hydrolysis, although this requirement is unusual for bacterial transcription initiation. (B) The interaction of NtrC and RNA polymerase, with the intervening DNA looped out, can be seen in the electron microscope. (B, courtesy of Harrison Echols and Sydney Kustu.)

Table 7–2 Sigma Factors of *E. coli*

SIGMA FACTOR	PROMOTERS RECOGNIZED
σ^{70}	most genes
σ^{32}	genes induced by heat shock
σ^{28}	genes for stationary phase and stress response
σ^{28}	genes involved in motility and chemotaxis
σ^{54}	genes for nitrogen metabolism
σ^{24}	genes dealing with misfolded proteins in the periplasm

The sigma factor designations refer to their approximate molecular weights, in kilodaltons.

RNA chain. Although this is one of the main ways in which both eucaryotes and procaryotes control transcription initiation, some bacteria and their viruses use an additional strategy based on interchangeable subunits of RNA polymerase. As described in Chapter 6, a sigma (σ) subunit is required for the bacterial RNA polymerase to recognize a promoter. Most bacteria produce a whole range of sigma subunits, each of which can interact with the RNA polymerase core and direct it to a different set of promoters (Table 7–2). This scheme permits one large set of genes to be turned off and a new set to be turned on simply by replacing one sigma subunit with another; the strategy is efficient because it bypasses the need to deal with genes one by one. Indeed, some bacteria code for nearly one hundred different sigma subunits and therefore rely heavily on this form of gene regulation. Bacterial viruses often use it subversively to take over the host polymerase and activate several sets of viral genes rapidly and sequentially (Figure 7–43).

Complex Switches Have Evolved to Control Gene Transcription in Eucaryotes

Bacteria and eucaryotes share many principles of gene regulation, including the key role played by gene regulatory proteins that bind tightly to short stretches of DNA, the importance of weak protein–protein actions in gene activation, and the versatility afforded by DNA looping. However, by comparison, gene regulation in eucaryotes involves many more proteins, much longer stretches of DNA, and often seems bewilderingly complex. This increased complexity provides the eucaryotic cell with an important advantage. Genetic switches in bacteria, as we have seen, typically respond to one or a few signals. But in eucaryotes it is common for dozens of signals to converge on a single promoter, with the transcription machinery integrating all these different signals to produce the appropriate level of mRNA. We begin our description of eucaryotic gene regulation by outlining the main features that distinguish it from gene regulation in bacteria.

- As discussed in Chapter 6, eucaryotic RNA polymerase II, which transcribes all the protein-coding genes, requires five general transcription factors (27 subunits *in toto*, see Table 6–3, p. 341), whereas bacterial RNA polymerase needs only a single general transcription factor, the σ subunit. As we have seen, the stepwise assembly of the general transcription factors at a eucaryotic promoter provides, in principle, multiple steps at which the

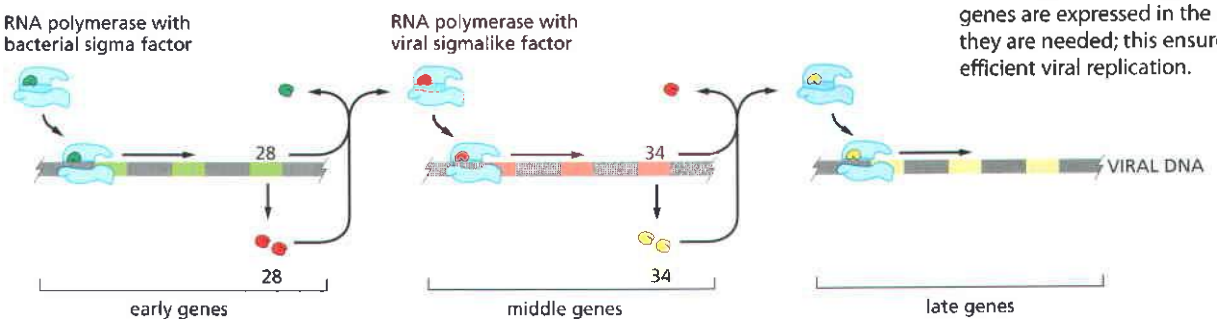


Figure 7–43 Interchangeable RNA polymerase subunits as a strategy to control gene expression in a bacterial virus. The bacterial virus SPO1, which infects the bacterium *B. subtilis*, uses the bacterial polymerase to transcribe its early genes immediately after the viral DNA enters the cell. One of the early genes, called 28, encodes a sigmalike factor that binds to RNA polymerase and displaces the bacterial sigma factor. This new form of polymerase specifically initiates transcription of the SPO1 "middle" genes. One of the middle genes encodes a second sigmalike factor, 34, that displaces the 28 product and directs RNA polymerase to transcribe the "late" genes. This last set of genes produces the proteins that package the virus chromosome into a virus coat and lyse the cell. By this strategy, sets of virus genes are expressed in the order in which they are needed; this ensures a rapid and efficient viral replication.

cell can speed up or slow down the rate of transcription initiation in response to gene regulatory proteins.

- Eucaryotic cells lack operons—sets of related genes transcribed as a unit—and therefore must regulate each gene individually.
- Each bacterial gene is typically controlled by one or only a few gene regulatory proteins, but it is common in eucaryotes for genes to be controlled by many (sometimes hundreds) of different regulatory proteins. This complexity is possible because, as we shall see, many eucaryotic gene regulatory proteins can act over very large distances (tens of thousands of nucleotide pairs) along DNA, allowing an almost unlimited number of them to influence the expression of a single gene.
- A central component of gene regulation in eucaryotes is *Mediator*, a 24-subunit complex, which serves as an intermediary between gene regulatory proteins and RNA polymerase (see Figure 6–19). Mediator provides an extended contact area for gene regulatory proteins compared to that provided by RNA polymerase alone, as in bacteria.
- The packaging of eucaryotic DNA into chromatin provides many opportunities for transcriptional regulation not available to bacteria.

Having discussed the general transcription factors for RNA polymerase II in Chapter 6 (see pp. 340–343), we focus here on the last four of these features and how they are used to control eucaryotic gene expression.

A Eucaryotic Gene Control Region Consists of a Promoter Plus Regulatory DNA Sequences

Because the typical eucaryotic gene regulatory protein controls transcription when bound to DNA far away from the promoter, the DNA sequences that control the expression of a gene are often spread over long stretches of DNA. We use the term **gene control region** to describe the whole expanse of DNA involved in regulating and initiating transcription of a gene, including the **promoter**, where the general transcription factors and the polymerase assemble, and all of the **regulatory sequences** to which gene regulatory proteins bind to control the rate of the assembly processes at the promoter (**Figure 7–44**). In animals and plants, it is not unusual to find the regulatory sequences of a gene dotted over distances as great as 50,000 nucleotide pairs. Much of this DNA serves as "spacer" sequences that gene regulatory proteins do not directly recognize, but this DNA may provide the flexibility needed for efficient DNA looping. In this context, it is important to remember that, like other regions of eucaryotic chromosomes, most of the DNA in gene control regions is packaged into nucleosomes and higher-order forms of chromatin, thereby compacting its length and altering its properties.

In this chapter, we shall loosely use the term **gene** to refer only to a segment of DNA that is transcribed into RNA (see Figure 7–44). However, the classical view of a gene includes the gene control region as well, making most eucaryotic genes considerably larger. The discovery of alternative RNA splicing has further complicated the definition of a gene—a point we discussed briefly in Chapter 6 and will return to later in this chapter.

It is the gene regulatory proteins that allow the genes of an organism to be turned on or off individually. In contrast to the small number of general transcription factors, which are abundant proteins that assemble on the promoters of all genes transcribed by RNA polymerase II, there are thousands of different gene regulatory proteins. For example, of the roughly 25,000 human genes, an estimated 8% (~2000 genes) encode gene regulatory proteins. Most of these recognize DNA sequences using one of the DNA-binding motifs described previously. Not surprisingly, the eucaryotic cell regulates each of its many genes in a unique way. Given the sheer number of genes in eucaryotes and the complexity of their regulation, it has been difficult to formulate simple rules for gene regulation that apply in every case. We can, however, make some generalizations about how gene regulatory proteins, once bound to gene control regions on DNA, set in motion the train of events that lead to gene activation or repression.

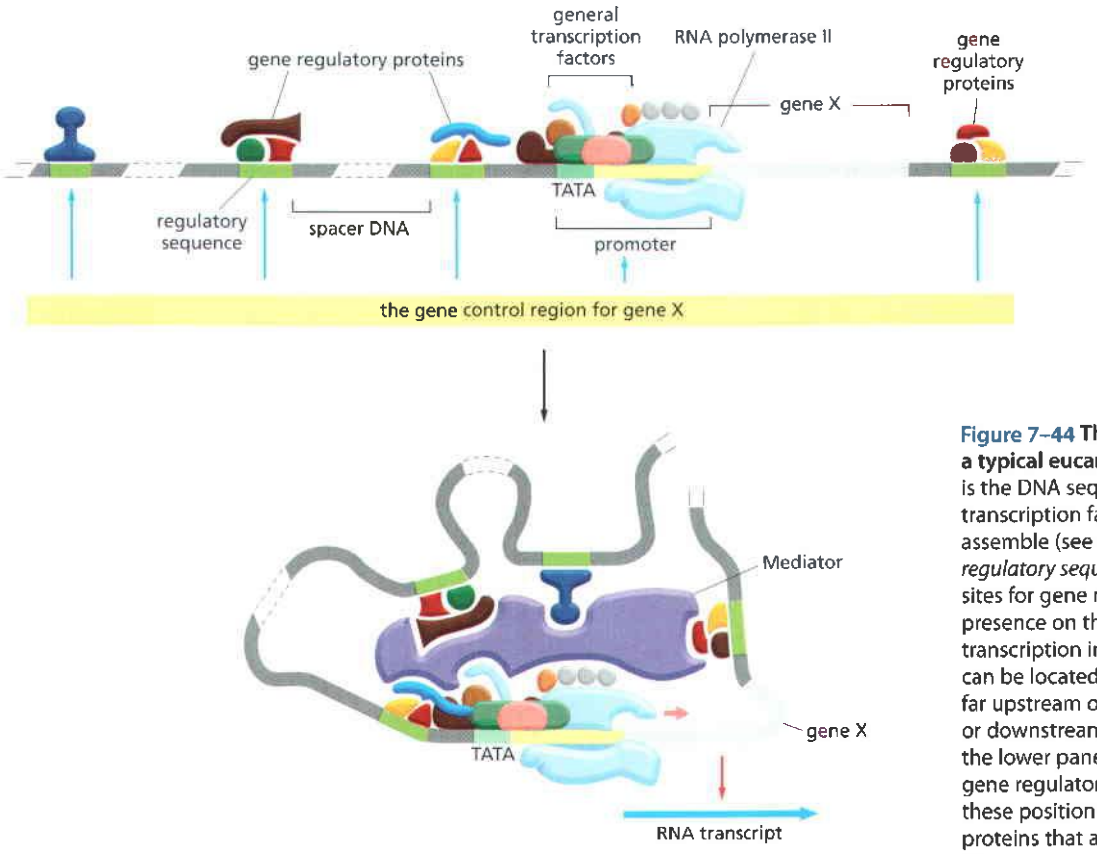


Figure 7–44 The gene control region for a typical eucaryotic gene. The *promoter* is the DNA sequence where the general transcription factors and the polymerase assemble (see Figure 6–16). The *regulatory sequences* serve as binding sites for gene regulatory proteins, whose presence on the DNA affects the rate of transcription initiation. These sequences can be located adjacent to the promoter, far upstream of it, or even within introns or downstream of the gene. As shown in the lower panel, DNA looping allows gene regulatory proteins bound at any of these positions to interact with the proteins that assemble at the promoter. Many gene regulatory proteins act through Mediator, while others influence the general transcription factors and RNA polymerase directly. Although not shown here, many gene regulatory proteins also influence the chromatin structure of the DNA control region thereby affecting transcription initiation indirectly (see Figure 4–45). As noted in the text, for simplicity, "gene X" refers here to the coding sequence within the gene.

Whereas Mediator and the general transcription factors are the same for all polymerase II transcribed genes, the gene regulatory proteins and the locations of their binding sites relative to the promoter differ for each gene.

Eucaryotic Gene Activator Proteins Promote the Assembly of RNA Polymerase and the General Transcription Factors at the Startpoint of Transcription

The DNA sites to which eucaryotic gene activator proteins bind were originally called *enhancers* because their presence "enhanced" the rate of transcription initiation. It came as a surprise when it was first discovered that these activator proteins could be bound tens of thousands of nucleotide pairs away from the promoter, but, as we have seen, DNA looping provides at least one explanation for this initially puzzling observation.

The simplest gene activator proteins have a modular design consisting of two distinct domains. One domain usually contains one of the structural motifs discussed previously that recognizes a specific DNA sequence. The second domain—sometimes called an *activation domain*—accelerates the rate of transcription initiation. This type of modular design was first revealed by experiments in which genetic engineering techniques were used to create a chimeric protein containing the activation domain of one protein fused to the DNA-binding domain of a different protein (**Figure 7–45**).

Once bound to DNA, how do eucaryotic gene activator proteins increase the rate of transcription initiation? As we will see shortly, there are several mechanisms by which this can occur, and, in many cases, these different mechanisms work in concert at a single promoter. But, regardless of the precise biochemical pathway, the ultimate function of activators is to attract, position, and modify the general transcription factors, Mediator, and RNA polymerase II at the promoter so that transcription can begin. They do this both by acting directly on these components and, indirectly, by changing the chromatin structure around the promoter.

Some activator proteins bind directly to one or more of the general transcription factors, accelerating their assembly on a promoter that is linked through DNA to that activator. Others interact with Mediator and attract it to DNA where it can then facilitate assembly of RNA polymerase and the general transcription factors at the promoter (see Figure 7–44). In this sense, eucaryotic

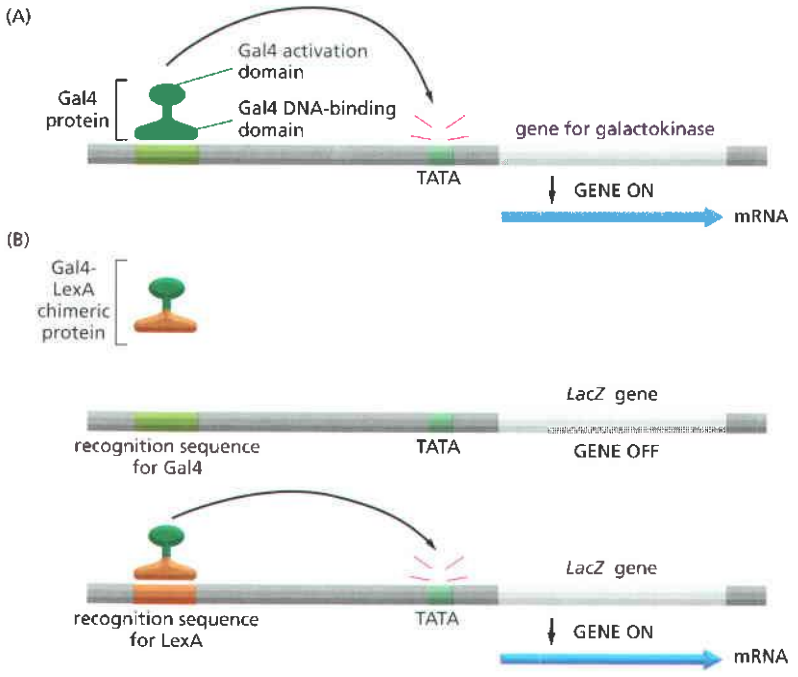


Figure 7–45 The modular structure of a gene activator protein. Outline of an experiment that reveals the presence of independent DNA-binding and transcription-activating domains in the yeast gene activator protein Gal4. A functional activator can be reconstituted from the C-terminal portion of the yeast Gal4 protein if it is attached to the DNA-binding domain of a bacterial gene regulatory protein (the LexA protein) by genetic engineering techniques. When the resulting bacterial–yeast hybrid protein is produced in yeast cells, it will activate transcription from yeast genes provided that the specific DNA-binding site for the bacterial protein has been inserted next to them. (A) Gal4 is normally responsible for activating the transcription of yeast genes that code for the enzymes that convert galactose to glucose. (B) A chimeric gene regulatory protein, produced by genetic engineering techniques, requires a LexA recognition sequence to activate transcription. In the experiment shown, the control region for one of the genes controlled by LexA was fused to the *E. coli* *LacZ* gene, which codes for the enzyme β-galactosidase (see Figure 7–39). β-Galactosidase is very simple to detect biochemically and thus provides a convenient way to monitor the expression level specified by a gene control region. As used here, *LacZ* is said to serve as a *reporter gene*, since it "reports" the activity of a gene control region.

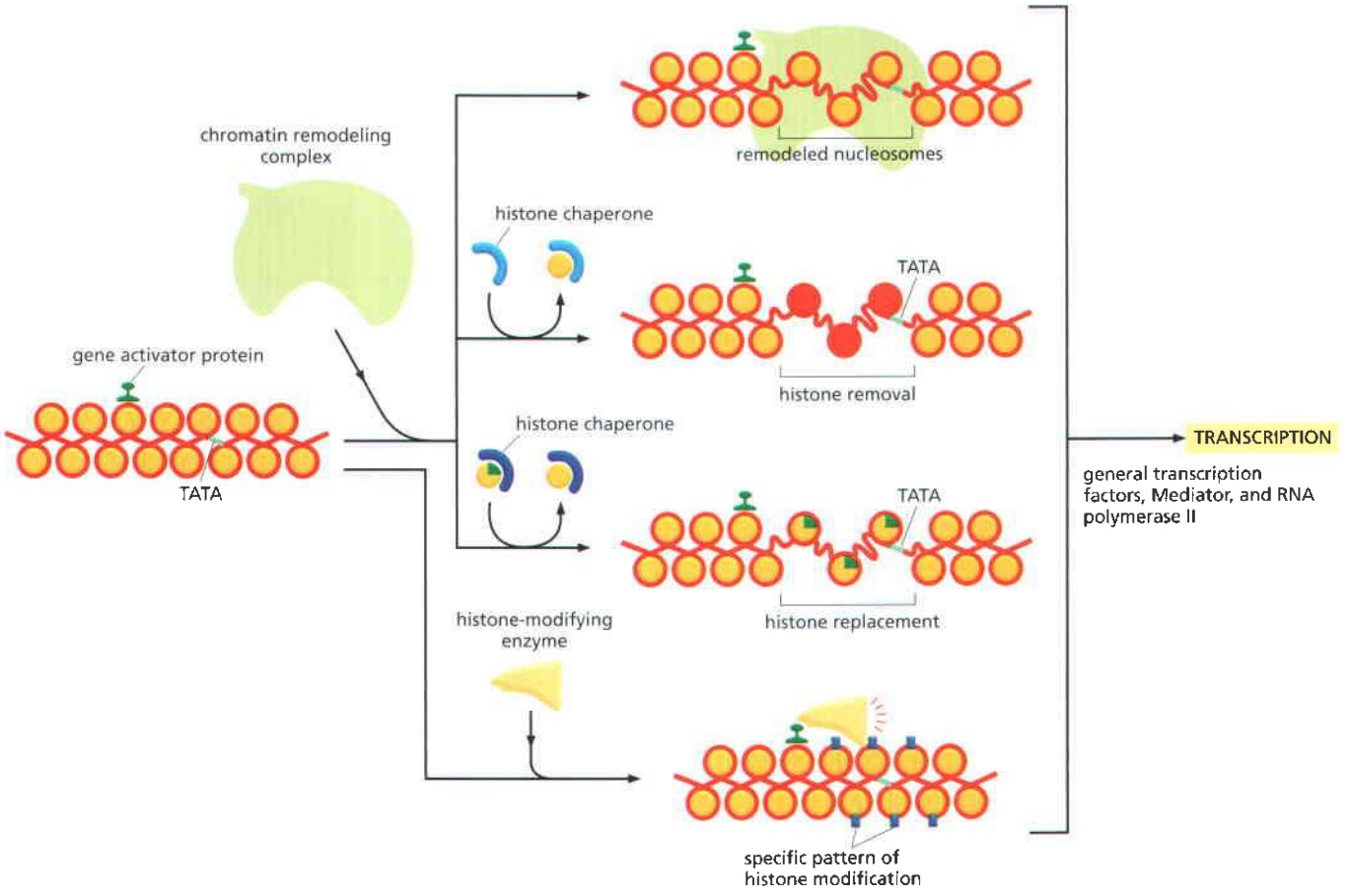
activators resemble those of bacteria in recruiting RNA polymerase to specific sites on DNA so it can begin transcribing.

Eucaryotic Gene Activator Proteins Also Modify Local Chromatin Structure

The general transcription factors, Mediator, and RNA polymerase seem unable on their own to assemble on a promoter that is packaged in standard nucleosomes. Indeed, it has been proposed such packaging may have evolved to prevent "leaky" transcription. In addition to their direct actions in assembling the transcription machinery at the promoter, gene activator proteins also promote transcription initiation by changing the chromatin structure of the regulatory sequences and promoters of genes.

As we saw in Chapter 4, four of the most important ways of locally altering chromatin are through covalent histone modifications, nucleosome remodeling, nucleosome removal, and nucleosome replacement. Gene activator proteins use all four of these mechanisms by attracting histone modification enzymes, ATP-dependent chromatin remodeling complexes, and histone chaperones to alter the chromatin structure of promoters they control (**Figure 7–46**). In general terms, these local alterations in chromatin structure are believed to make the underlying DNA more accessible, thereby facilitating the assembly of the general transcription factors, Mediator, and RNA polymerase at the promoter. Local chromatin modification also allows additional gene regulatory proteins to bind to the control region of the gene. However, the most important role of covalent histone modifications in transcription is probably not in directly changing chromatin structure; rather, as discussed in Chapter 4, these modifications provide favorable interactions for the binding of a large set of proteins that read a "histone code." For transcription initiation, these proteins include other histone-modifying enzymes (reader–writer complexes), chromatin remodeling complexes, and at least one of the general transcription factors (**Figure 7–47**).

The alterations of chromatin structure that occur during transcription initiation can persist for variable lengths of time. In some cases, as soon as the gene regulatory protein dissociates from DNA, the chromatin modifications are rapidly reversed, restoring the gene to its pre-activated state. This rapid reversal is especially important for genes that the cell must quickly switch on and off in



response to external signals, such as the glucocorticoid hormone discussed earlier in this chapter. However, in other cases, the altered chromatin structure seems to persist, even after the gene regulatory protein that directed its establishment has dissociated from DNA. In principle, this memory can extend into the next cell generation because, as discussed in Chapter 4, chromatin structure can be self-renewing (see Figure 4–52). It is interesting to consider the possibility that different histone modifications persist for different times in order to provide the cell with a mechanism for long-, medium-, and short-term memory of gene expression patterns.

A special type of chromatin modification occurs as RNA polymerase II transcribes through a gene. In most cases, the nucleosomes just ahead of the polymerase are acetylated by writer complexes carried by the polymerase, removed by histone chaperones, and deposited behind the moving polymerase. They are then rapidly deacetylated and methylated, also by reader–writer complexes that are carried by the polymerase, leaving behind nucleosomes that are especially resistant to transcription. Although this remarkable process may seem counterintuitive, it likely evolved to prevent spurious transcription re-initiation behind a moving polymerase, which is, in essence, clearing a path through chromatin. Later in this chapter, when we discuss *RNA interference*, the potential dangers to the cell of such inappropriate transcription will become especially obvious.

We have just seen that gene activator proteins can profoundly influence chromatin structure. However, even before these activator proteins are brought into play, many genes are "poised" to become rapidly activated. For example, the regulatory regions for many genes are "marked" by a short, nucleosome-free region flanked by nucleosomes that contain the histone variant H2AZ. This arrangement, which is specified by DNA sequence, allows free access of gene regulatory proteins to the nucleosome-free region; in addition, the H2AZ-containing nucleosomes are thought to be easily disassembled, thus further facilitating transcription initiation.

Figure 7–46 Four ways eucaryotic activator proteins can direct local alterations in chromatin structure to stimulate transcription initiation. Although shown as separate pathways, these mechanisms often work together during the activation of a gene. For example, prior acetylation of histones makes it easier for histone chaperones to remove them from nucleosomes. A few patterns of histone modification that promote transcription initiation are listed in Figure 4–44, and a specific example is given in Figure 7–47. Nucleosome remodeling and histone removal favor transcription initiation by increasing the accessibility of DNA and thereby facilitating the binding of Mediator, RNA polymerase, and the general transcription factors as well as additional activator proteins. Transcription initiation and the formation of a compact chromatin structure can be regarded as competing biochemical assembly reactions, and enzymes that increase—even transiently—the accessibility of DNA in chromatin will tend to favor transcription initiation.

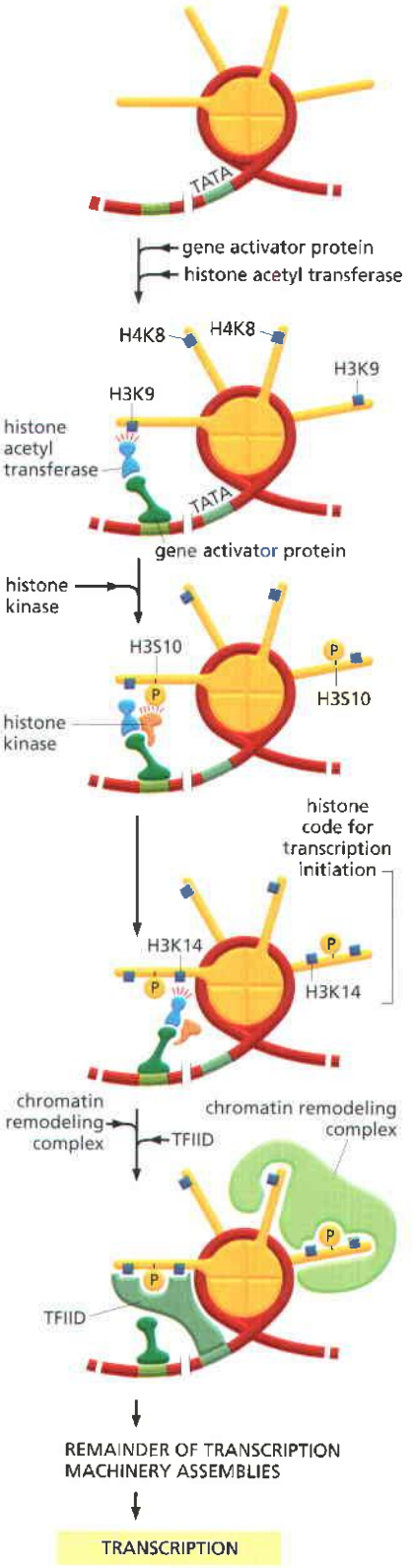


Figure 7–47 Writing and reading the histone code during transcription initiation. In this example, taken from the human interferon gene promoter, a gene activator protein binds to DNA packaged into chromatin and first attracts a histone acetyl transferase to acetylate lysine 9 of histone H3 and lysine 8 of histone H4. Next, a histone kinase, attracted by the gene activator protein, phosphorylates serine 10 of histone H3, but can only do so after lysine 9 has been acetylated. The serine modification then signals the histone acetyl transferase to acetylate position K14 of histone H3. At this point the histone code for transcription initiation, set into motion by the binding of the gene activator protein, has been written. Note that the writing is sequential, with each histone modification depending on a prior modification.

The final reading of the code occurs when the general transcription factor TFIID and the chromatin remodeling complex SWI/SNF bind, both of which strongly promote the subsequent steps of transcription initiation. TFIID and SWI/SNF both recognize acetylated histone tails through a *bromodomain*, a protein domain specialized to read this particular mark on histones; a bromodomain is carried in a subunit of each protein complex. (Adapted from T. Agalioti, G. Chen and D. Thanos, *Cell* 111:381–392, 2002. With permission from Elsevier.)

Gene Activator Proteins Work Synergistically

We have seen that eucaryotic gene activator proteins can influence different steps in transcription initiation. In general, where several factors work together to enhance a reaction rate, the joint effect is not merely the sum of the enhancements that each factor alone contributes, but the product. If, for example, factor A lowers the free-energy barrier for a reaction by a certain amount and thereby speeds up the reaction 100-fold, and factor B, by acting on another aspect of the reaction, does likewise, then A and B acting in parallel will lower the barrier by a double amount and speed up the reaction 10,000-fold. Even if A and B work simply by attracting the same protein, the affinity of that protein for the reaction site increases multiplicatively. Thus, gene activator proteins often exhibit *transcriptional synergy*, where several activator proteins working together produce a transcription rate that is much higher than that of the sum of the activators working alone (**Figure 7–48**). It is not difficult to see how multiple gene regulatory proteins, each binding to a different regulatory DNA sequence, work together to control the final rate of transcription of a eucaryotic gene.

Since gene activator proteins can influence many different steps on the pathway to transcriptional activation, it is worth considering whether these steps always occur in a prescribed order. For example, does histone modification always precede chromatin remodeling, as in the example of Figure 7–47? Does Mediator enter before or after RNA polymerase? The answers to these questions appear to be different for different genes—and even for the same gene under the influence of different gene regulatory proteins (**Figure 7–49**).



Figure 7–48 Transcriptional synergy. This experiment compares the rate of transcription produced by three experimentally constructed regulatory regions in a eucaryotic cell and reveals transcriptional synergy, the greater than additive effect of multiple activators. Transcriptional synergy is typically observed between different gene activator proteins from the same organism and even between activator proteins from different eucaryotic species when they are experimentally introduced into the same cell. This last observation reflects the high degree of conservation of the elaborate machinery responsible for eucaryotic transcription initiation.

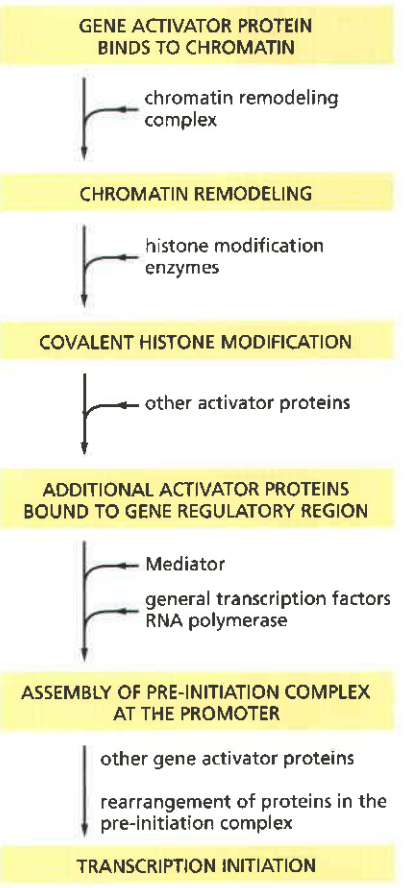


Figure 7–49 An order of events leading to transcription initiation of a specific gene. In this well-studied example from the budding yeast *S. cerevisiae*, the steps toward transcription initiation occur in a particular order; however, this order differs from one gene to the next. For example, at another gene, histone modification occurs first, followed by RNA polymerase recruitment, followed by chromatin remodeling complex recruitment. Figure 7–47 illustrates yet another possible order of events.

Whatever the precise mechanisms and the order in which they are carried out, a gene regulatory protein must be bound to DNA either directly or indirectly to influence transcription of its target gene, and the rate of transcription of a gene ultimately depends upon the spectrum of regulatory proteins bound upstream and downstream of its transcription start site.

Eucaryotic Gene Repressor Proteins Can Inhibit Transcription in Various Ways

Like bacteria, eucaryotes use **gene repressor proteins** in addition to activator proteins to regulate transcription of their genes. However, because of differences in the way that eucaryotes and bacteria initiate transcription, eucaryotic repressors have many more possible mechanisms of action. We saw in Chapter 4 that large regions of the genome can be shut down by the packaging of DNA into heterochromatin. However, eucaryotic genes are rarely organized along the genome according to function, so this strategy is not generally useful for most examples of gene regulation. Instead, most eucaryotic repressors must work on a gene-by-gene basis. Unlike bacterial repressors, most eucaryotic repressors do not directly compete with the RNA polymerase for access to the DNA; rather they use a variety of other mechanisms, some of which are illustrated in **Figure 7–50**. Like gene activator proteins, many eucaryotic repressor proteins act through more than one mechanism at a given target gene, thereby ensuring robust and efficient repression.

Gene repression is especially important to animals and plants whose growth depends on elaborate and complex developmental programs. Misexpression of a single gene at a critical time can have disastrous consequences for the individual. For this reason, many of the genes encoding the most important developmental regulatory proteins are kept tightly repressed when they are not needed.

Eucaryotic Gene Regulatory Proteins Often Bind DNA Cooperatively

So far we have seen that when eucaryotic activator and repressor proteins bind to specific DNA sequences, they set in motion a complex series of events that culminate in transcription initiation or its opposite, repression. However, these proteins rarely recognize DNA as individual polypeptides. In reality, efficient DNA binding in the eucaryotic cell typically requires several sequence-specific DNA proteins acting together. For example, two gene regulatory proteins with a weak affinity for each other might cooperate to bind to a DNA sequence, neither protein having a sufficient affinity for DNA to bind to the DNA site on its own. In one well-studied case, the DNA-bound protein dimer creates a distinct surface that is recognized by a third protein that carries an activator domain that stimulates transcription. This example illustrates an important general point: protein–protein interactions that are too weak to form complexes in solution can do so on DNA, with the DNA sequence acting as a "crystallization" site or seed for the assembly of a protein complex.

As shown in **Figure 7–51**, an individual gene regulatory protein can often participate in more than one type of regulatory complex. A protein might function, for example, in one case as part of a complex that activates transcription and in another case as part of a complex that represses transcription. Thus,

individual eucaryotic gene regulatory proteins are not necessarily dedicated activators or repressors; instead, they function as regulatory parts that are used to build complexes whose function depends on the final assembly of all of the individual components. This final assembly, in turn, depends both on the arrangement of control region DNA sequences and on the particular gene regulatory proteins present in active form in the cell. Each eucaryotic gene is therefore regulated by a "committee" of proteins, all of which must be present to express the gene at its proper level.

In some cases, the precise DNA sequence to which a regulatory protein binds directly can affect the conformation of this protein and thereby influence its subsequent transcriptional activity. When bound to one type of DNA sequence, for example, a steroid hormone receptor protein interacts with a co-repressor and ultimately turns off transcription. When bound to a slightly different DNA sequence, it assumes a different conformation and interacts with a coactivator, thereby stimulating transcription.

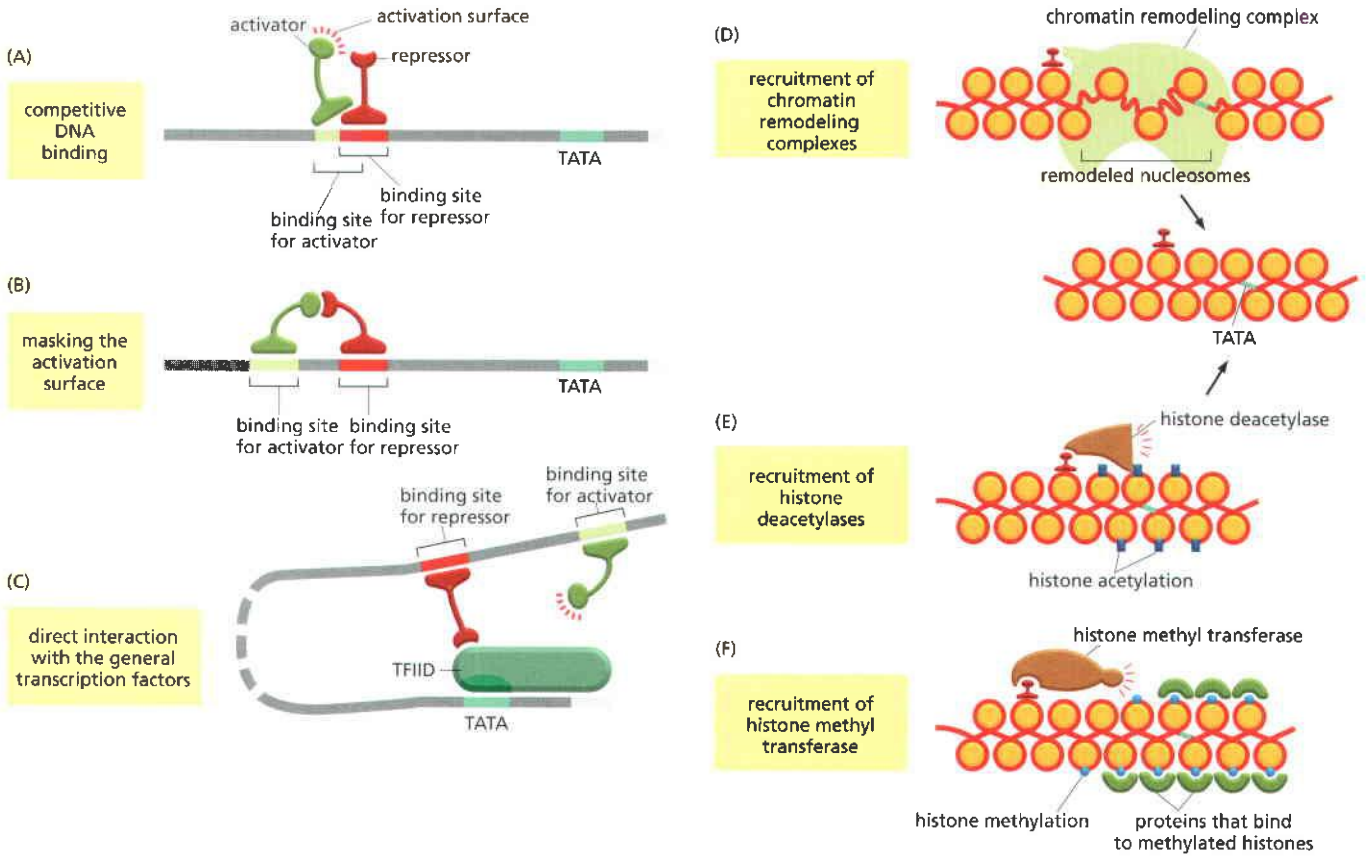


Figure 7–50 Six ways in which eucaryotic gene repressor proteins can operate. (A) Gene activator proteins and gene repressor proteins compete for binding to the same regulatory DNA sequence. (B) Both proteins can bind DNA, but the repressor binds to the activation domain of the activator protein, thereby preventing it from carrying out its activation functions. In a variation of this strategy, the repressor binds tightly to the activator without having to be bound to DNA directly. (C) The repressor blocks assembly of the general transcription factors. Some repressors also act at late stages in transcription initiation, for example, by preventing the release of the RNA polymerase from the general transcription factors. (D) The repressor recruits a chromatin remodeling complex which returns the nucleosomal state of the promoter region to its pre-transcriptional form. (E) The repressor attracts a histone deacetylase to the promoter. As we have seen, histone acetylation can stimulate transcription initiation (Figure 7–47), and the repressor simply reverses this modification. (F) The repressor attracts a histone methyl transferase which modifies certain positions on histones which, in turn, are bound by proteins that maintain the chromatin in a transcriptionally silent form. For example, in *Drosophila*, the histone methyl transferase Suv39 methylates the K9 position of histone H3, a modification that is bound by the HP1 protein. In another example, E(z) methylates the K27 position of H3, and this modification is bound by the Polycomb protein. HP1 and Polycomb recognize methylated lysines through a *chromodomain*. They can act locally to turn off specific genes, as shown here, or can occupy a whole region of a chromosome to repress a cluster of genes. A seventh mechanism of negative control—inactivation of a transcriptional activator by heterodimerization—is illustrated in Figure 7–24. For simplicity, nucleosomes have been omitted from (A)–(C), and the scales of (D)–(F) have been reduced relative to (A)–(C).

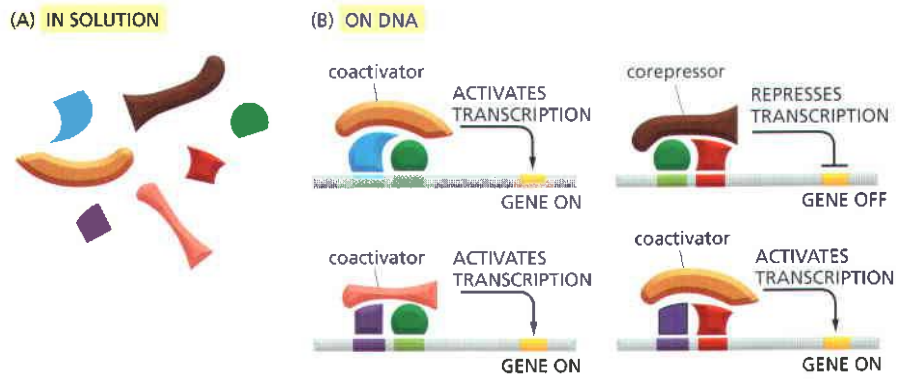


Figure 7–51 Eucaryotic gene regulatory proteins often assemble into complexes on DNA. Seven gene regulatory proteins are shown in (A). The nature and function of the complex they form depends on the specific DNA sequence that seeds their assembly. In (B), some assembled complexes activate gene transcription, while another represses transcription. Note that both the *red* and the *green* proteins are shared by both activating and repressing complexes. Proteins that do not themselves bind DNA but assemble on other DNA-bound gene regulatory proteins are often termed coactivators or co-repressors. However, these terms are somewhat confusing because they encompass an enormous variety of proteins including histone readers and writers, chromatin remodeling complexes, and many other classes of proteins. Some have no intrinsic activity themselves but simply serve as a "scaffolding" to attract those that do.

Typically, a few relatively short stretches of nucleotide sequence guide the assembly of a group of regulatory proteins on DNA (see Figure 7–51). However, in some extreme cases of regulation by committee a more elaborate protein–DNA structure is formed (**Figure 7–52**). Since the final assembly requires the presence of many gene regulatory proteins that bind DNA, it provides a simple way to ensure that a gene is expressed only when the cell contains the correct combination of these proteins. We saw earlier how the formation of heterodimers in solution provides a mechanism for the combinatorial control of gene expression. The assembly of complexes of gene regulatory proteins on DNA provides a second important mechanism for combinatorial control, one that offers far richer opportunities.

Complex Genetic Switches That Regulate *Drosophila* Development Are Built Up from Smaller Modules

Given that gene regulatory proteins can be positioned at multiple sites along long stretches of DNA, that these proteins can assemble into complexes at each site, and that the complexes influence the chromatin structure as well as the recruitment and assembly of the general transcription machinery at the promoter, there would seem to be almost limitless possibilities for the elaboration of control devices to regulate eucaryotic gene transcription.

A particularly striking example of a complex, multicomponent genetic switch is that controlling the transcription of the *Drosophila Even-skipped* (*Eve*) gene, whose expression plays an important part in the development of the *Drosophila* embryo. If this gene is inactivated by mutation, many parts of the embryo fail to form, and the embryo dies early in development. As discussed in Chapter 22, at the stage of development when *Eve* begins to be expressed, the embryo is a single giant cell containing multiple nuclei in a common cytoplasm. This cytoplasm is not uniform, however: it contains a mixture of gene regulatory proteins that are distributed unevenly along the length of the embryo, thus providing positional information that distinguishes one part of the embryo from another (**Figure 7–53**). (The way these differences are initially set up is discussed in Chapter 22.) Although the nuclei are initially identical, they rapidly begin to express different genes because they are exposed to different gene regulatory proteins. The nuclei near the anterior end of the developing embryo, for example, are exposed to a set of gene regulatory proteins that is distinct from the set that influences nuclei at the posterior end of the embryo.

The regulatory DNA sequences controlling the *Eve* gene are designed to read the concentrations of gene regulatory proteins at each position along the length of the embryo and to interpret this information in such a way that the *Eve* gene is expressed in seven stripes, each initially five to six nuclei wide and positioned precisely along the anterior–posterior axis of the embryo (**Figure 7–54**). How is this remarkable feat of information processing carried out? Although not all of the molecular details are understood, several general principles have emerged from studies of *Eve* and other *Drosophila* genes that are similarly regulated.

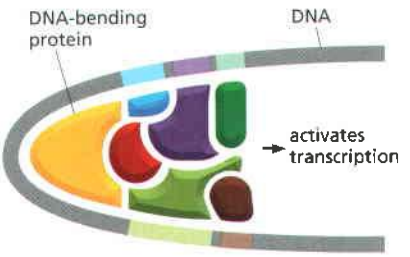


Figure 7–52 Schematic depiction of a committee of gene regulatory proteins bound to an enhancer. The protein shown in *yellow* is called an architectural protein since its main role is to bend the DNA to allow the cooperative assembly of the other components. The structure depicted here is based on that found in the control region of the gene that codes for a subunit of the T cell receptor (discussed in Chapter 25), and it activates transcription at a nearby promoter. Only certain cells of the developing immune system, which eventually give rise to mature T cells, have the complete set of proteins needed to form this structure.

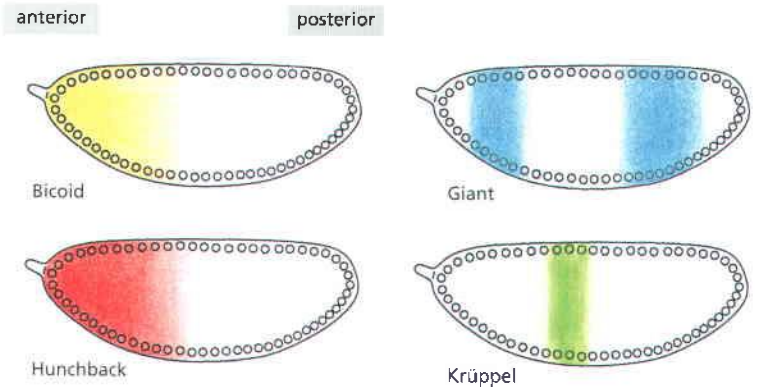


Figure 7–53 The nonuniform distribution of four gene regulatory proteins in an early *Drosophila* embryo. At this stage the embryo is a syncytium, with multiple nuclei in a common cytoplasm. Although the detail is not shown in these drawings, all of these proteins are concentrated in the nuclei.

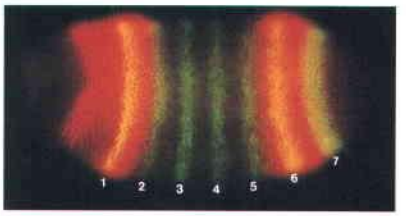


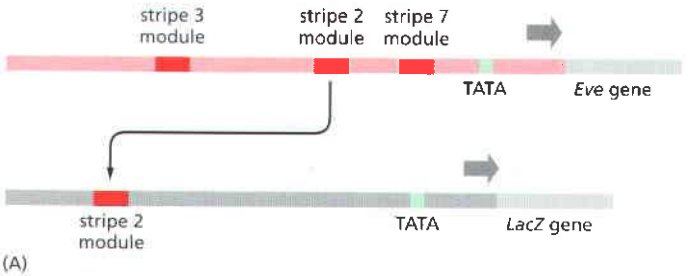
Figure 7–54 The seven stripes of the protein encoded by the *Even-skipped (Eve)* gene in a developing *Drosophila* embryo. Two and one-half hours after fertilization, the egg was fixed and stained with antibodies that recognize the Eve protein *(green)* and antibodies that recognize the Giant protein *(red)*. Where Eve and Giant proteins are both present, the staining appears *yellow*. At this stage in development, the egg contains approximately 4000 nuclei. The Eve and Giant proteins are both located in the nuclei, and the Eve stripes are about four nuclei wide. The staining pattern of the Giant protein is also shown in Figure m7–52/7–53. (Courtesy of Michael Levine.)

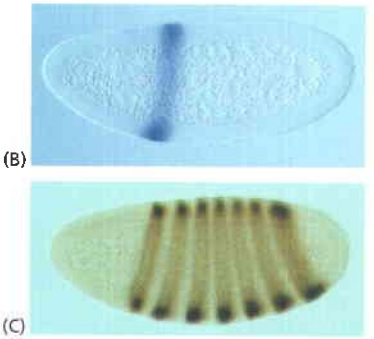
The regulatory region of the *Eve* gene is very large (approximately 20,000 nucleotide pairs). It is formed from a series of relatively simple regulatory modules, each of which contains multiple regulatory sequences and is responsible for specifying a particular stripe of *Eve* expression along the embryo. This modular organization of the *Eve* gene control region is revealed by experiments in which a particular regulatory module (say, that specifying stripe 2) is removed from its normal setting upstream of the *Eve* gene, placed in front of a reporter gene (see Figure 7–45), and reintroduced into the *Drosophila* genome. When developing embryos derived from flies carrying this genetic construct are examined, the reporter gene is found to be expressed in precisely the position of stripe 2 (**Figure 7–55**). Similar experiments reveal the existence of other regulatory modules, each of which specifies either one of the other six stripes or some other part of the *Eve* expression pattern normally displayed at later stages of development (see Figure 22–39).

The *Drosophila Eve* Gene Is Regulated by Combinatorial Controls

A detailed study of the stripe 2 regulatory module has provided insights into how it reads and interprets positional information. It contains recognition sequences for two gene regulatory proteins (Bicoid and Hunchback) that activate *Eve* transcription and two (Krüppel and Giant) that repress it (**Figure 7–56**). (The gene regulatory proteins of *Drosophila* often have colorful names reflecting the phenotype that results if the gene encoding the protein is inactivated by mutation.) The relative concentrations of these four proteins determine whether the protein complexes that form at the stripe 2 module activate transcription of the *Eve* gene. **Figure 7–57** shows the distributions of the four gene regulatory proteins across the region of a *Drosophila* embryo where stripe 2 forms. It is thought that either of the two repressor proteins, when bound to the DNA, will turn off the

Figure 7–55 Experiment demonstrating the modular construction of the *Eve* gene regulatory region. (A) A 480-nucleotide-pair piece of the Eve regulatory region was removed and inserted upstream of a test promoter that directs the synthesis of the enzyme β-galactosidase (the product of the *E. coli LacZ* gene). (B) When this artificial construct was reintroduced into the genome of *Drosophila* embryos, the embryos expressed β-galactosidase (detectable by histochemical staining) precisely in the position of the second of the seven Eve stripes (C). (B and C, courtesy of Stephen Small and Michael Levine.)





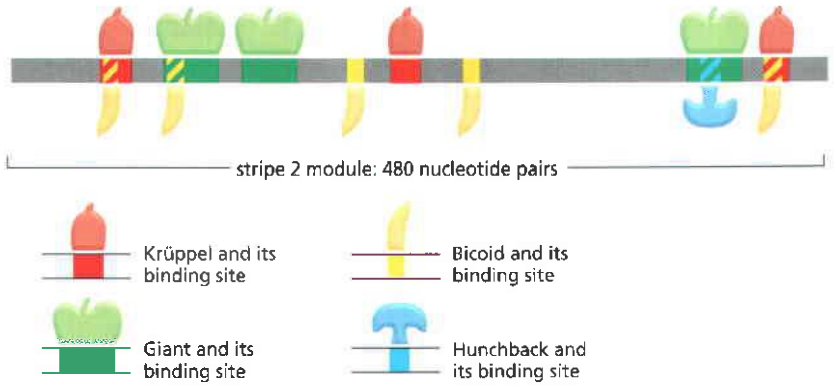


Figure 7–56 Close-up view of the *Eve* stripe 2 unit. The segment of the *Eve* gene control region identified in the previous figure contains regulatory sequences, each of which binds one or another of four gene regulatory proteins. It is known from genetic experiments that these four regulatory proteins are responsible for the proper expression of *Eve* in stripe 2. Flies that are deficient in the two gene activators Bicoid and Hunchback, for example, fail to express efficiently *Eve* in stripe 2. In flies deficient in either of the two gene repressors, Giant and Krüppel, stripe 2 expands and covers an abnormally broad region of the embryo. The DNA binding sites for these gene regulatory proteins were determined by cloning the genes encoding the proteins, overexpressing the proteins in *E. coli*, purifying them, and performing DNA-footprinting experiments (see Figure 7–29). The *top* diagram indicates that, in some cases, the binding sites for the gene regulatory proteins overlap and the proteins can compete for binding to the DNA. For example, binding of Krüppel and binding of Bicoid to the site at the far right are thought to be mutually exclusive.

stripe 2 module, whereas both Bicoid and Hunchback must bind for its maximal activation. This simple regulatory unit thereby combines these four positional signals so as to turn on the stripe 2 module (and therefore the expression of the *Eve* gene) only in those nuclei that are located where the levels of both Bicoid and Hunchback are high and both Krüppel and Giant are absent. This combination of activators and repressors occurs in only one region of the early embryo; everywhere else, therefore, the stripe 2 module is silent.

We have thus far discussed two mechanisms of combinatorial control of gene expression—heterodimerization of gene regulatory proteins in solution (see Figure 7–20) and the assembly of combinations of gene regulatory proteins into small complexes on DNA (see Figure 7–51). It is likely that both mechanisms participate in the complex regulation of *Eve* expression. In addition, the regulation of stripe 2 just described illustrates a third type of combinatorial control.

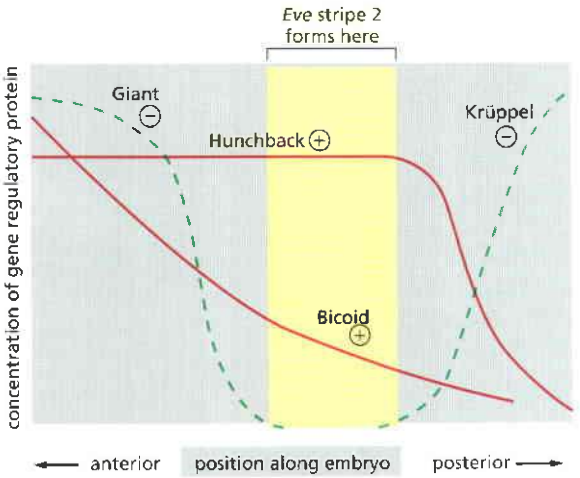


Figure 7–57 Distribution of the gene regulatory proteins responsible for ensuring that *Eve* is expressed in stripe 2. The distributions of these proteins were visualized by staining a developing *Drosophila* embryo with antibodies directed against each of the four proteins (see Figures 7–53 and 7–54). The expression of *Eve* in stripe 2 occurs only at the position where the two activators (Bicoid and Hunchback) are present and the two repressors (Giant and Krüppel) are absent. In fly embryos that lack Krüppel, for example, stripe 2 expands posteriorly. Likewise, stripe 2 expands posteriorly if the DNA-binding sites for Krüppel in the stripe 2 module (see Figure 7–56) are inactivated by mutation.

The *Eve* gene itself encodes a gene regulatory protein, which, after its pattern of expression is set up in seven stripes, regulates the expression of other *Drosophila* genes. As development proceeds, the embryo is thus subdivided into finer and finer regions that eventually give rise to the different body parts of the adult fly, as discussed in Chapter 22.

This example from *Drosophila* embryos is unusual in that the nuclei are exposed directly to positional cues in the form of concentrations of gene regulatory proteins. In embryos of most other organisms, individual nuclei are in separate cells, and extracellular positional information must either pass across the plasma membrane or, more usually, generate signals in the cytosol in order to influence the genome.

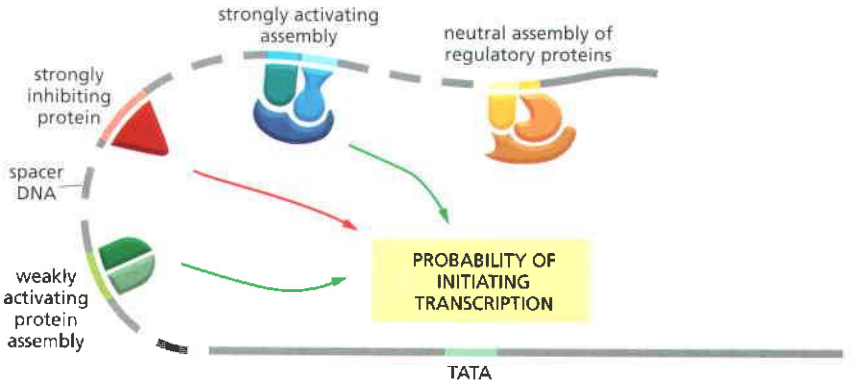


Figure 7–58 The integration of multiple inputs at a promoter. Multiple sets of gene regulatory proteins can work together to influence transcription initiation at a promoter, as they do in the *Eve* stripe 2 module illustrated in Figure 7–56. It is not yet understood in detail how the cell achieves integration of multiple inputs, but it is likely that the final transcriptional activity of the gene results from a competition between activators and repressors that act by the mechanisms summarized in Figures 7–46 and 7–50.

Because the individual regulatory sequences in the *Eve* stripe 2 module are strung out along the DNA, many sets of gene regulatory proteins can be bound simultaneously at separate sites and influence the promoter of a gene. The promoter integrates the transcriptional cues provided by all of the bound proteins (**Figure 7–58**).

The regulation of *Eve* expression is an impressive example of combinatorial control. Seven combinations of gene regulatory proteins—one combination for each stripe—activate *Eve* expression, while many other combinations (all those found in the interstripe regions of the embryo) keep the stripe elements silent. The other stripe regulatory modules are thought to be constructed similarly to those described for stripe 2, being designed to read positional information provided by other combinations of gene regulatory proteins. The entire gene control region, strung out over 20,000 nucleotide pairs of DNA, binds more than 20 different regulatory proteins. A large and complex control region is thereby built from a series of smaller modules, each of which consists of a unique arrangement of short DNA sequences recognized by specific gene regulatory proteins.

Complex Mammalian Gene Control Regions Are Also Constructed from Simple Regulatory Modules

Perhaps 8% of the coding capacity of a mammalian genome is devoted to the synthesis of proteins that serve as regulators of gene transcription. This large number of genes reflects the exceedingly complex network of controls governing expression of mammalian genes. Each gene is regulated by a set of gene regulatory proteins; each of those proteins is the product of a gene that is in turn regulated by a whole set of other proteins, and so on. Moreover, the regulatory protein molecules are themselves influenced by signals from outside the cell, which can make them active or inactive in a whole variety of ways (**Figure 7–59**). Thus, we can view the pattern of gene expression in a cell as the result of a complicated molecular computation that the intracellular gene control network performs in response to information from the cell's surroundings. We shall discuss these issues further in Chapters 15 and 22, which deal with cell signaling and development, but the complexity is remarkable even at the level of an individual genetic switch regulating the activity of a single gene. It is not unusual, for example, to find a mammalian gene with a control region that is 100,000 nucleotide pairs in length, in which many modules, each containing a number of regulatory sequences that bind gene regulatory proteins, are interspersed with long stretches of other noncoding DNA.

One of the best-understood examples of a complex mammalian regulatory region is found in the human β-globin gene, which is expressed exclusively in red blood cells. A complex array of gene regulatory proteins controls the expression of the gene, some acting as activators and others as repressors (**Figure 7–60**). The concentrations (or activities) of many of these gene regulatory proteins change during development, and only a particular combination of all the proteins triggers transcription of the gene. The human β-globin gene is part of a cluster of globin genes (**Figure 7–61**A) that are all transcribed exclusively in erythroid cells,

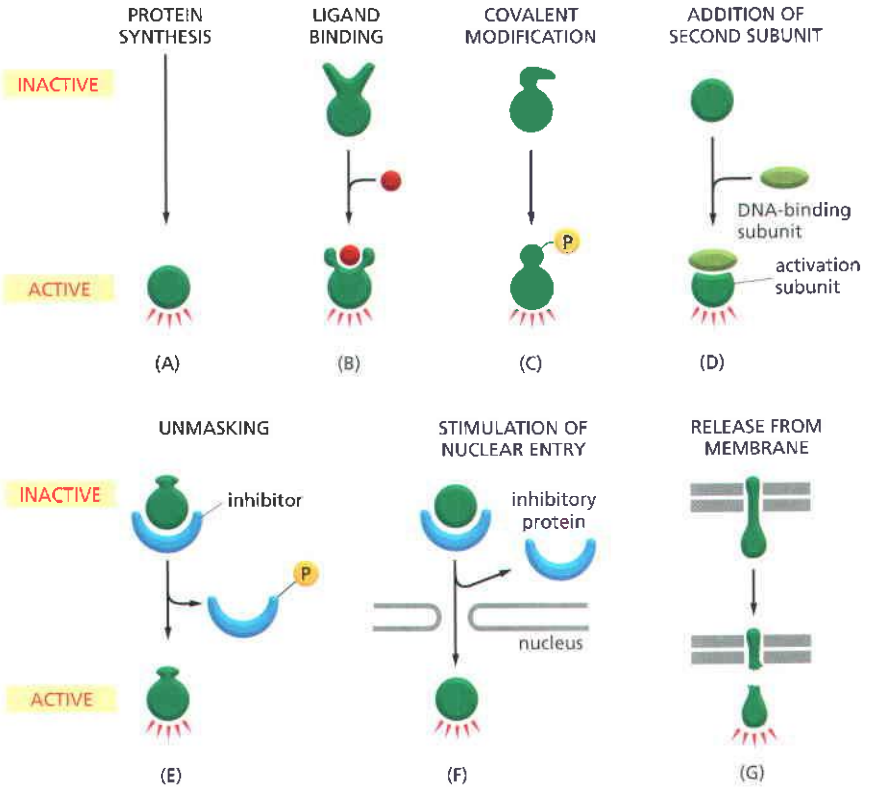


Figure 7–59 Some ways in which the activity of gene regulatory proteins is regulated in eucaryotic cells. (A) The protein is synthesized only when needed and is rapidly degraded by proteolysis so that it does not accumulate. (B) Activation by ligand binding. (C) Activation by covalent modification. Phosphorylation is shown here, but many other modifications are possible (see Table 3–3, p. 186). (D) Formation of a complex between a DNA-binding protein and a separate protein with a transcription-activating domain. (E) Unmasking of an activation domain by the phosphorylation of an inhibitor protein. (F) Stimulation of nuclear entry by removal of an inhibitory protein that otherwise keeps the regulatory protein from entering the nucleus. (G) Release of a gene regulatory protein from a membrane bilayer by regulated proteolysis.

that is, cells of the red blood cell lineage, but at different stages of mammalian development (see Figure 7–61B). The ε-globin gene is expressed in the early embryo, γ in the later embryo and fetus, and δ and β primarily in the adult. The gene products differ slightly in their oxygen-binding properties, suiting them for the different oxygenation conditions in the embryo, fetus, and adult. Each of the globin genes has its own set of regulatory proteins that are necessary to turn the gene on at the appropriate time.

The globin genes are unusual in that, at the appropriate time and place, they are transcribed at extremely high rates: indeed, red blood cells are little more

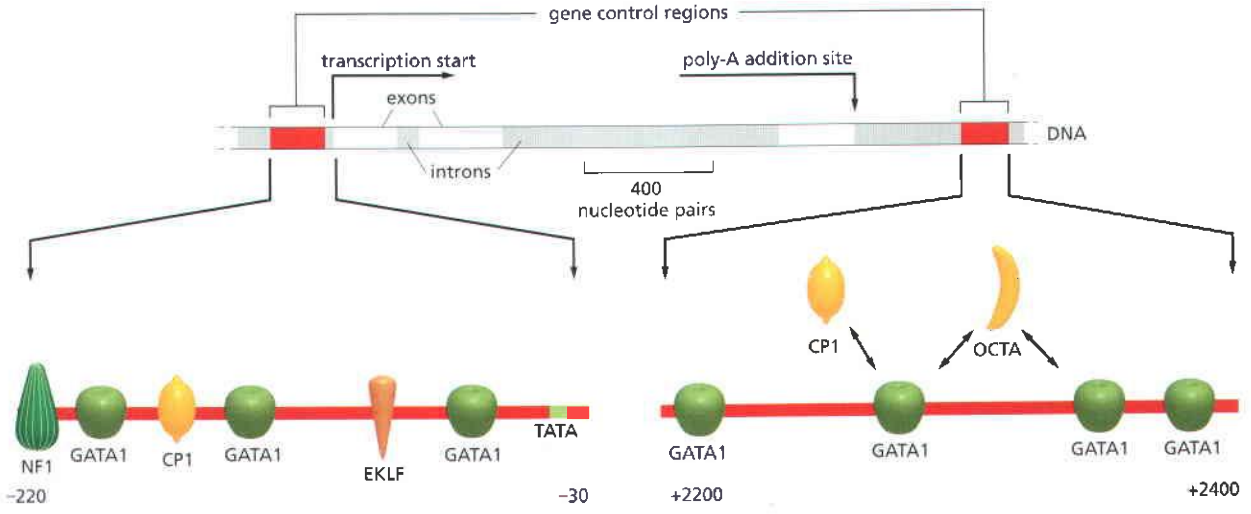


Figure 7–60 Model for the control of the human β-globin gene. The diagram shows some of the gene regulatory proteins that control expression of the gene during red blood cell development (see Figure 7–61). Some of the gene regulatory proteins shown, such as CP1, are found in many types of cells, while others, such as GATA1, are present in only a few types of cells—including red blood cells—and therefore are thought to contribute to the cell-type specificity of β-globin gene expression. As indicated by the *double-headed arrows*, several of the binding sites for GATA1 overlap those of other gene regulatory proteins; it is thought that by binding to these sites, GATA1 excludes binding of other proteins. Once bound to DNA, the gene regulatory proteins recruit chromatin remodeling complexes, histone modifying enzymes, the general transcription factors, Mediator, and RNA polymerase to the promoter. (Adapted from B. Emerson, in Gene Expression: General and Cell-Type Specific [M. Karin, ed.], pp. 116–161. Boston: Birkhauser, 1993.)

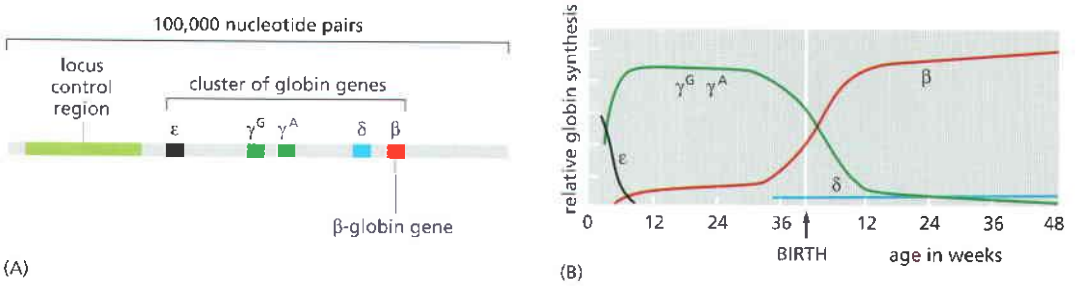


Figure 7–61 The cluster of β-like globin genes in humans. (A) The large chromosomal region shown spans 100,000 nucleotide pairs and contains the five globin genes and a locus control region (LCR). (B) Changes in the expression of the β-like globin genes at various stages of human development. Each of the globin chains encoded by these genes combines with an α-globin chain to form the hemoglobin in red blood cells (see Figure 4–86). (A, after F. Grosveld, G.B. van Assendelft, D.R. Greaves and G. Kollias, *Cell* 51:975–985, 1987. With permission from Elsevier.)

than bags of hemoglobin that was synthesized by precursor cells. To achieve this extraordinarily high level of transcription, the globin genes, in addition to their individual regulatory sequences, share a control region called the *locus control region* (LCR), which lies far upstream from the gene cluster and is needed for the proper expression of each gene in the cluster (see Figure 7–61A). The importance of the LCR can be seen in patients with a certain type of thalassemia, a severe inherited form of anemia. In these patients, the β-globin locus has suffered a deletion that removes all or part of the LCR. Although the β-globin and its nearby regulatory region are intact, the gene remains transcriptionally silent, even in erythroid cells.

The way in which the LCR functions is not understood in detail, but it is known that the gene regulatory proteins that bind the LCR interact, through DNA looping, with proteins bound to the control regions of the globin genes they regulate. In this way, the proteins bound at the LCR help attract chromatin remodeling complexes, histone-modifying enzymes, and components of the transcription machinery that act in conjunction with the specific regulatory regions of each individual globin gene. In addition, the LCR includes a *barrier sequence* (see Figure 4–47) that prevents the spread of neighboring heterochromatin into the β-globin locus, as discussed in Chapter 4. This dual feature distinguishes the globin LCR from many other types of regulatory sequences in the human genome; however, the globin genes are not alone in having an LCR, as LCRs are also present upstream of other highly transcribed, cell-type-specific genes. We should probably think of LCRs, not as unique DNA elements with specialized properties, but rather as especially powerful combinations of more fundamental types of regulatory sequences.

Insulators Are DNA Sequences That Prevent Eucaryotic Gene Regulatory Proteins from Influencing Distant Genes

All genes have control regions, which dictate at which times, under what conditions, and in what tissues the gene will be expressed. We have also seen that eucaryotic gene regulatory proteins can act across very long stretches of DNA. How, then, are control regions of different genes kept from interfering with one another? In other words, what keeps a gene regulatory protein bound on the control region of one gene from inappropriately influencing the transcription of adjacent genes?

To avoid such cross-talk, several types of DNA elements function to compartmentalize the genome into discrete regulatory domains. In Chapter 4 we discussed barrier sequences that prevent the spread of heterochromatin into genes that need to be expressed. A second type of DNA element, called an insulator, prevents enhancers from running amok and activating inappropriate genes (**Figure 7–62**). An insulator can apparently block the communication

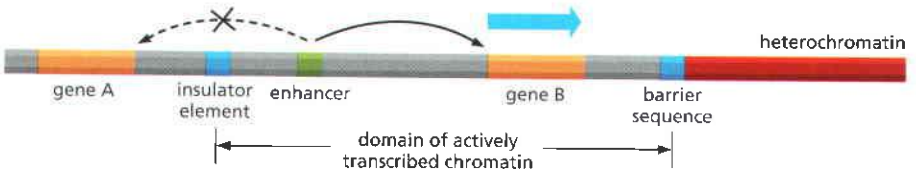


Figure 7–62 Schematic diagram summarizing the properties of insulators and barrier sequences. Insulators directionally block the action of enhancers *(left-hand side)*, and barrier sequences prevent the spread of heterochromatin *(right-hand side)*. Thus gene B is properly regulated and gene B's enhancer is prevented from influencing the transcription of gene A. How barrier sequences are likely to function is depicted in Figure 4–47. It is not yet understood how insulators exert their effects; one possibility is that they serve as "decoys," tying up the transcriptional machinery and preventing it from interacting with an authentic enhancer. Another is that they anchor DNA to the nuclear envelope, thereby interfering with DNA looping between an enhancer and an inappropriate promoter.

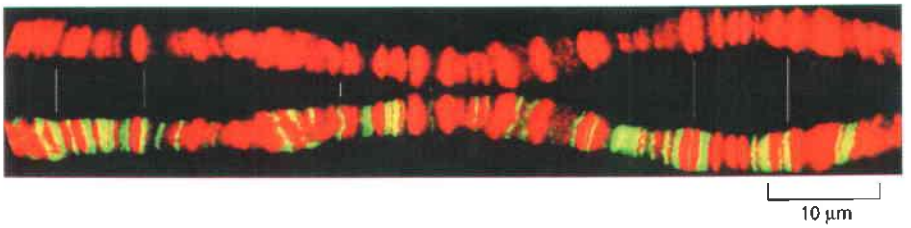


Figure 7–63 Localization of a *Drosophila* insulator-binding protein on polytene chromosomes. A polytene chromosome (discussed in Chapter 4) was stained with propidium iodide *(red)* to show its banding patterns—with bands appearing *bright red* and interbands as dark gaps in the pattern *(top)*. The positions on this polytene chromosome that are bound by a particular insulator protein (called BEAF) are stained *bright green* using antibodies directed against the protein *(bottom)*. BEAF is preferentially localized to interband regions, reflecting its role in organizing chromosomes into structural, as well as functional, domains. For convenience, these two micrographs of the same polytene chromosome are arranged as mirror images. (Courtesy of Uli Laemmli, from K. Zhao et al., *Cell* 81:879–889, 1995. With permission from Elsevier.)

between an enhancer and a promoter, but, to do so, it must be located between the two. Although proteins that bind to insulators have been identified, how they directionally neutralize enhancers is still a mystery.

Even though their mechanisms are not understood in detail, the distribution of insulators and barrier sequences in a genome is thought to divide it into independent domains of gene regulation and chromatin structure (**Figure 7–63**). Aspects of this organization can be visualized by staining whole chromosomes for the specialized proteins that bind these DNA elements.

Although chromosomes are organized into orderly domains that discourage enhancers from acting indiscriminately, there are special circumstances where an enhancer located one chromosome has been found to activate a gene located on a second chromosome. A remarkable example occurs in the regulation of the mammalian olfactory receptors. These are the proteins expressed by sensory neurons that allow mammals to discriminate accurately among many thousands of distinct smells (see p. 917). Humans, for example, have 350 olfactory receptor genes, and they are carefully regulated so that only one of these genes is expressed in each sensory neuron. The olfactory receptor genes are dispersed among many different chromosomes, but there is only a single enhancer for all of them. Once this enhancer activates a receptor gene by associating with its regulatory region, it remains stably associated thereby precluding activation of any of the other receptor genes. Although there is much we do not understand about this mechanism, it does indicate the extreme versatility of transcriptional regulation strategies.

Gene Switches Rapidly Evolve

We have seen that the control regions of eucaryotic genes are often spread out over long stretches of DNA, whereas those of procaryotic genes are typically clustered around the start point of transcription. It seems likely that the close-packed arrangement of bacterial genetic switches developed from more extended forms of switches in response to the evolutionary pressure on bacteria to maintain a small genome size. This compression comes at a price, however, as it restricts the complexity and adaptability of the control device. In contrast, the extended form of eucaryotic control regions—with discrete regulatory modules separated by long stretches of spacer DNA—facilitates the reshuffling of regulatory modules during evolution, both to create new regulatory circuits and to modify old ones. As we saw in Chapters 1 and 4, and we shall see again in Chapter 22, changes in gene regulation—rather than the acquisition of new genes—underlie much of the wide variety of life on Earth. Unraveling the history of how modern gene control regions have evolved presents a fascinating challenge to biologists, with many clues available in present-day genomes.

Summary

Gene regulatory proteins switch the transcription of individual genes on and off in cells. In procaryotes these proteins usually bind to specific DNA sequences close to the RNA polymerase start site and, depending on the nature of the regulatory protein and the precise location of its binding site relative to the start site, either activate or repress transcription of the gene. The flexibility of the DNA helix, however, also allows proteins bound at distant sites to affect the RNA polymerase at the promoter by the looping out of the intervening DNA. The regulation of higher eucaryotic genes is much more complex, commensurate with a larger genome size and the large variety of cell types that

are formed. A single eucaryotic gene is typically controlled by many gene regulatory proteins bound to sequences that can be thousands of nucleotide pairs from the promoter that directs transcription of the gene. Eucaryotic activators and repressors act by a wide variety of mechanisms—generally altering chromatin structure and controlling the assembly of the general transcription factors, Mediator, and RNA polymerase at the promoter. The time and place that each gene is transcribed, as well as its rates of transcription under different conditions, are determined by the spectrum of gene regulatory proteins that bind to the regulatory region of the gene.

THE MOLECULAR GENETIC MECHANISMS THAT CREATE SPECIALIZED CELL TYPES

Although all cells must be able to switch genes on and off in response to changes in their environments, the cells of multicellular organisms have evolved this capacity to an extreme degree and in highly specialized ways to form an organized array of differentiated cell types. In particular, once a cell in a multicellular organism becomes committed to differentiate into a specific cell type, the cell maintains this choice through many subsequent cell generations, which means that it remembers the changes in gene expression involved in the choice. This phenomenon of *cell memory* is a prerequisite for the creation of organized tissues and for the maintenance of stably differentiated cell types. In contrast, other changes in gene expression in eucaryotes, as well as most in bacteria, are only transient. The tryptophan repressor, for example, switches off the tryptophan genes in bacteria only in the presence of tryptophan; as soon as tryptophan is removed from the medium, the genes are switched back on, and the descendants of the cell will have no memory that their ancestors had been exposed to tryptophan. Even in bacteria, however, a few types of changes in gene expression can be inherited stably.

In this section we shall examine not only cell memory mechanisms, but also how gene regulatory devices can be combined to create "logic circuits" through which cells integrate signals, keep time, remember events in their past, and adjust the levels of gene expression over entire chromosomes. We begin by considering some of the best-understood genetic mechanisms of cell differentiation, which operate in bacterial and yeast cells.

DNA Rearrangements Mediate Phase Variation in Bacteria

We have seen that cell differentiation in higher eucaryotes usually occurs without detectable changes in DNA sequence. In some procaryotes, in contrast, a stably inherited pattern of gene regulation is achieved by DNA rearrangements that activate or inactivate specific genes. Since a change in DNA sequence will be copied faithfully during all subsequent DNA replication cycles, an altered state of gene activity will be inherited by all the progeny of the cell in which the rearrangement occurred. Some of these DNA rearrangements are, however, reversible so that occasional individuals can switch back to the original DNA configuration. The result is an alternating pattern of gene activity that can be detected by observations over long time periods and many generations.

A well-studied example of this differentiation mechanism occurs in *Salmonella* bacteria and is known as **phase variation**. Although this mode of differentiation has no known counterpart in higher eucaryotes, it can have considerable impact on animals because disease-causing bacteria use it to evade detection by the immune system. The switch in *Salmonella* gene expression is brought about by the occasional inversion of a specific 1000-nucleotide-pair piece of DNA. This change alters the expression of the cell-surface protein flagellin, for which the bacterium has two different genes (**Figure 7–64**). A site-specific recombination enzyme catalyzes the inversion and thereby changes the orientation of a promoter that is located within the inverted DNA segment. With the promoter in one orientation, the bacteria synthesize one type of flagellin:

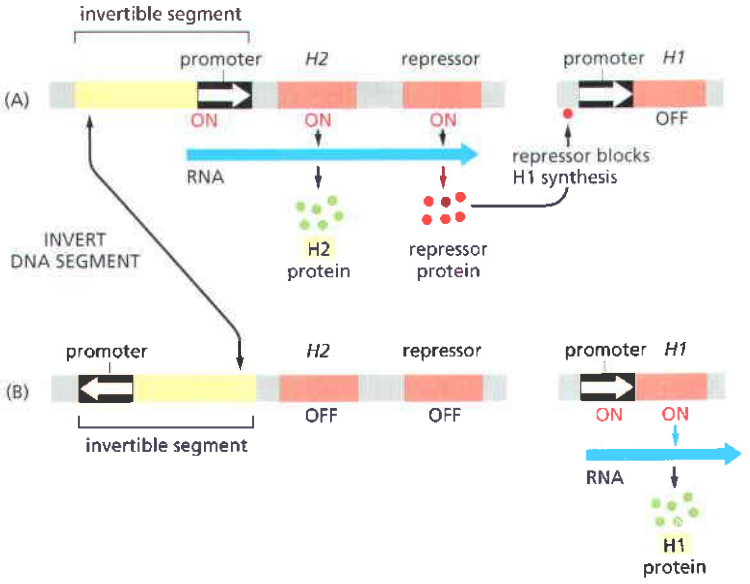


Figure 7–64 Switching gene expression by DNA inversion in bacteria. Alternating transcription of two flagellin genes in a *Salmonella* bacterium is caused by a simple site-specific recombination event that inverts a small DNA segment containing a promoter. (A) In one orientation, the promoter activates transcription of the *H2* flagellin gene as well as that of a repressor protein that blocks the expression of the *H1* flagellin gene. (B) When the promoter is inverted, it no longer turns on *H2* or the repressor, and the *H1* gene, which is thereby released from repression, is expressed instead. The recombination mechanism is activated only rarely (about once in every 10^5 cell divisions). Therefore, the production of one or other flagellin tends to be faithfully inherited in each clone of cells.

with the promoter in the other orientation, they synthesize the other type. Because inversions occur only rarely, entire clones of bacteria will have one type of flagellin or the other.

Phase variation almost certainly evolved because it protects the bacterial population against the immune response of its vertebrate host. If the host makes antibodies against one type of flagellin, a few bacteria whose flagellin has been altered by gene inversion will still be able to survive and multiply.

Bacteria isolated from the wild very often exhibit phase variation for one or more phenotypic traits. Standard laboratory strains of bacteria lose these "instabilities" over time, and underlying mechanisms have been studied in only a few cases. Not all involve DNA inversion. A bacterium that causes a common sexually transmitted human disease *(Neisseria gonorrhoeae)*, for example, avoids immune attack by means of a heritable change in its surface properties that arises from gene conversion (discussed in Chapter 5) rather than by inversion. This mechanism transfers DNA sequences from a library of silent "gene cassettes" to a site in the genome where the genes are expressed; it has the advantage of creating many variants of the major bacterial surface protein.

A Set of Gene Regulatory Proteins Determines Cell Type in a Budding Yeast

Because they are so easy to grow and to manipulate genetically, yeasts have served as model organisms for studying the mechanisms of gene control in eucaryotic cells. The common baker's yeast, *Saccharomyces cerevisiae*, has attracted special interest because of its ability to differentiate into three distinct cell types. *S. cerevisiae* is a single-celled eucaryote that exists in either a haploid or a diploid state. Diploid cells form by a process known as **mating**, in which two haploid cells fuse. In order for two haploid cells to mate, they must differ in *mating type* (sex). In *S. cerevisiae* there are two mating types, α and **a**, which are specialized for mating with each other. Each produces a specific diffusible signaling molecule (mating factor) and a specific cell-surface receptor protein. These jointly enable a cell to recognize and be recognized by its opposite cell type, with which it then fuses. The resulting diploid cells, called **a**/α, are distinct from either parent: they are unable to mate but can form spores (sporulate) when they run out of food, giving rise to haploid cells by the process of meiosis (discussed in Chapter 21).

The mechanisms by which these three cell types are established and maintained illustrate several of the strategies we have discussed for changing the pattern of gene expression. The mating type of the haploid cell is determined by a

single locus, the **mating-type (Mat) locus**, which in an **a**-type cell encodes a single gene regulatory protein, Mata1, and in an α cell encodes two gene regulatory proteins, Matα1 and Matα2. The Mata1 protein has no effect in the **a**-type haploid cell that produces it, but becomes important later in the diploid cell that results from mating. In contrast, the Matα2 protein acts in the α cell as a transcriptional repressor that turns off the **a**-specific genes, while the Matα1 protein acts as a transcriptional activator that turns on the α-specific genes. Once cells of the two mating types have fused, the combination of the Mata1 and Matα2 regulatory proteins generates a completely new pattern of gene expression, unlike that of either parent cell. **Figure 7–65** illustrates the mechanism by which the mating-type-specific genes are expressed in different patterns in the three cell types. This was among the first examples of combinatorial gene control to be identified, and it remains one of the best understood at the molecular level.

Although in most laboratory strains of *S. cerevisiae*, the **a** and α cell types are stably maintained through many cell divisions, some strains isolated from the wild can switch repeatedly between the **a** and α cell types by a mechanism of gene rearrangement whose effects are reminiscent of the DNA rearrangements in *N. gonorrhoeae*, although the exact mechanism seems to be peculiar to yeast. On either side of the *Mat* locus in the yeast chromosome, there is a silent locus encoding the mating-type gene regulatory proteins: the silent locus on one side encodes Matα1 and Matα2; the silent locus on the other side encodes Mata1. In approximately every other cell division, the active gene in the *Mat* locus is excised and replaced by a newly synthesized copy of the silent locus determining the opposite mating type. Because the change removes one gene from the active "slot" and replaces it by another, this mechanism is called the *cassette mechanism*. The change is reversible because, although the original gene at the *Mat* locus is discarded, a silent copy remains in the genome. New DNA copies made from the silent genes function as disposable cassettes that will be inserted in alternation into the *Mat* locus, which serves as the "playing head" (**Figure 7–66**).

The silent cassettes are packaged into a specialized form of chromatin and maintained in a transcriptionally inactive form. The study of these cassettes—

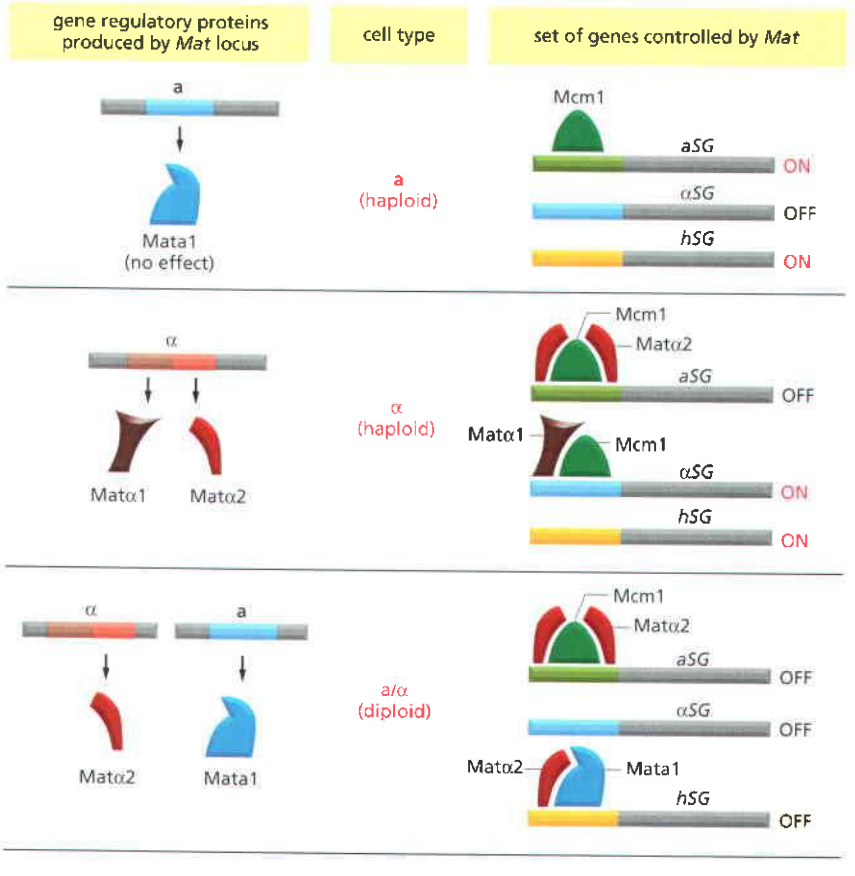


Figure 7–65 Control of cell type in yeasts. Three gene regulatory proteins (Matα1, Matα2, and Mata1) produced by the *Mat* locus determine yeast cell type. Different sets of genes are transcribed in haploid cells of type a, in haploid cells of type α, and in diploid cells (type a/α). The haploid cells express a set of haploid-specific genes *(hSG)* and either a set of α-specific genes *(αSG)* or a set of a-specific genes *(aSG)*. The diploid cells express none of these genes. The Mat regulatory proteins control many target genes in each type of cell by binding, in various combinations, to specific regulatory sequences upstream of these genes. Note that the Matα1 protein is a gene activator protein, whereas the Matα2 protein is a gene repressor protein. Both work in combination with a gene regulatory protein called Mcm1 that is present in all three cell types. In the diploid cell type, Matα2 and Mata1 form a heterodimer (shown in detail in Figure 7–21) that turns off a set of genes (including the gene encoding the Matα1 activator protein) different from that turned off by the Matα2 and Mcm1 proteins. This relatively simple system of gene regulatory proteins is an example of combinatorial control of gene expression.

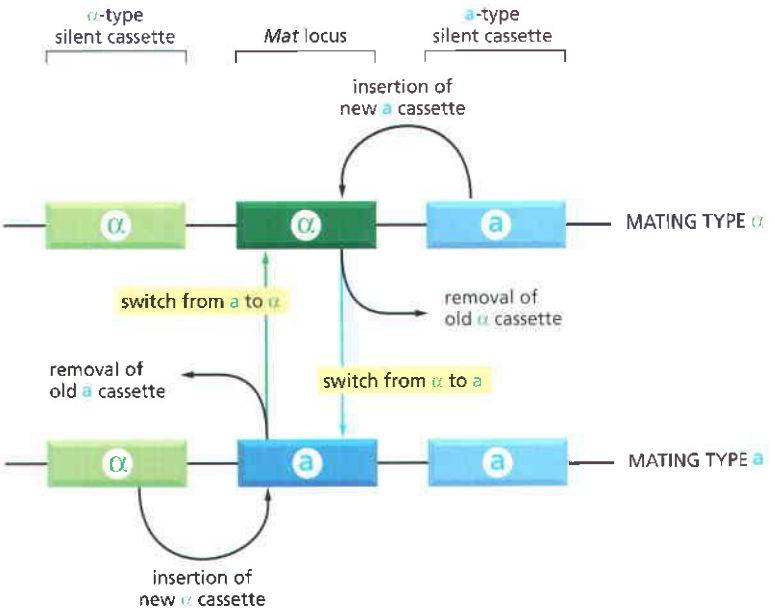


Figure 7–66 Cassette model of yeast mating-type switching. Cassette switching occurs by a gene-conversion process that involves a specialized enzyme (the HO endonuclease) that makes a double-stranded cut at a specific DNA sequence in the *Mat* locus. The DNA near the cut is then excised and replaced by a copy of the silent cassette of opposite mating type. The mechanism of this specialized form of gene conversion is similar to the repair of double-stranded breaks discussed in Chapter 5 (pp. 308–309).

which has been ongoing for nearly 40 years—has provided many of the key insights into the role of chromatin structure in gene regulation.

Two Proteins That Repress Each Other's Synthesis Determine the Heritable State of Bacteriophage Lambda

As we saw at the beginning of the present chapter, the nucleus of a single differentiated cell contains all the genetic information needed to construct a whole vertebrate or plant. This observation eliminates the possibility that an irreversible change in DNA sequence is a major mechanism of cell differentiation in these higher eucaryotes although such changes do occur in lymphocyte differentiation (discussed in Chapter 25). *Reversible* DNA sequence changes, resembling those just described for *Salmonella* and yeasts, in principle could still be responsible for some of the inherited changes in gene expression observed in higher organisms, but there is currently no evidence that such mechanisms are widely used.

Other mechanisms that we have already touched upon in this chapter, however, are also capable of producing patterns of gene regulation that can be inherited by subsequent cell generations. One of the simplest examples is found in the bacterial virus (bacteriophage) lambda where a switch causes the virus to flip-flop between two stable self-maintaining states. This type of switch can be viewed as a prototype for similar, but more complex, switches that operate in the development of higher eucaryotes.

We mentioned earlier that bacteriophage lambda can in favorable conditions become integrated into the *E. coli* cell DNA, to be replicated automatically each time the bacterium divides. Alternatively, the virus can multiply in the cytoplasm, killing its host (see Figure 5–78). Proteins encoded by the bacteriophage genome mediate the switch between these two states. The genome contains a total of about 50 genes, which are transcribed in very different patterns in the two states. A virus destined to integrate, for example, must produce the lambda *integrase* protein, which is needed to insert the lambda DNA into the bacterial chromosome, but must repress the production of the viral proteins responsible for virus multiplication. Once one transcriptional pattern or the other has been established, it is stably maintained.

At the heart of this complex gene regulatory switching mechanism are two gene regulatory proteins synthesized by the virus: the **lambda repressor protein** (cI protein), which we have already encountered, and the **Cro protein**. These proteins repress each other's synthesis, an arrangement that gives rise to just two stable states (**Figure 7–67**). In state 1 (the *prophage state)* the lambda



Figure 7–67 A simplified version of the regulatory system that determines the mode of growth of bacteriophage lambda in the *E. coli* host cell. In stable state 1 (the prophage state) the bacteriophage synthesizes a repressor protein, which activates its own synthesis and turns off the synthesis of several other bacteriophage proteins, including the Cro protein. In state 2 (the lytic state) the bacteriophage synthesizes the Cro protein, which turns off the synthesis of the repressor protein, so that many bacteriophage proteins are made and the viral DNA replicates freely in the *E. coli* cell, eventually producing many new bacteriophage particles and killing the cell. This example shows how two gene regulatory proteins can be combined in a circuit to produce two heritable states. Both the lambda repressor and the Cro protein recognize the operator through a helix–turn–helix motif (see Figure 7–11).

repressor occupies the operator, blocking the synthesis of Cro and also activating its own synthesis. In state 2 (the *lytic state)* the Cro protein occupies a different site in the operator, blocking the synthesis of repressor but allowing its own synthesis. In the prophage state most of the DNA of the stably integrated bacteriophage is not transcribed; in the lytic state, this DNA is extensively transcribed, replicated, packaged into new bacteriophage, and released by host cell lysis.

When the host bacteria are growing well, an infecting virus tends to adopt state 1, allowing the DNA of the virus to multiply along with the host chromosome. When the host cell is damaged, an integrated virus converts from state 1 to state 2 in order to multiply in the cell cytoplasm and make a quick exit. This conversion is triggered by the host response to DNA damage, which inactivates the repressor protein. In the absence of such interference, however, the lambda repressor both turns off production of the Cro protein and turns on its own synthesis, and this *positive feedback loop* helps to maintain the prophage state.

Simple Gene Regulatory Circuits Can Be Used to Make Memory Devices

Positive feedback loops provide a simple general strategy for cell memory—that is, for the establishment and maintenance of heritable patterns of gene transcription. **Figure 7–68** shows the basic principle, stripped to its barest essentials. Eucaryotic cells use many variations of this simple strategy. Several gene regulatory proteins that are involved in establishing the *Drosophila* body plan (discussed in Chapter 22), for example, stimulate their own transcription, thereby creating a positive feedback loop that promotes their continued synthesis; at the same time many of these proteins repress the transcription of genes encoding other important gene regulatory proteins. In this way, a few gene regulatory proteins that reciprocally affect one another's synthesis and activities can specify a sophisticated pattern of inherited behavior.

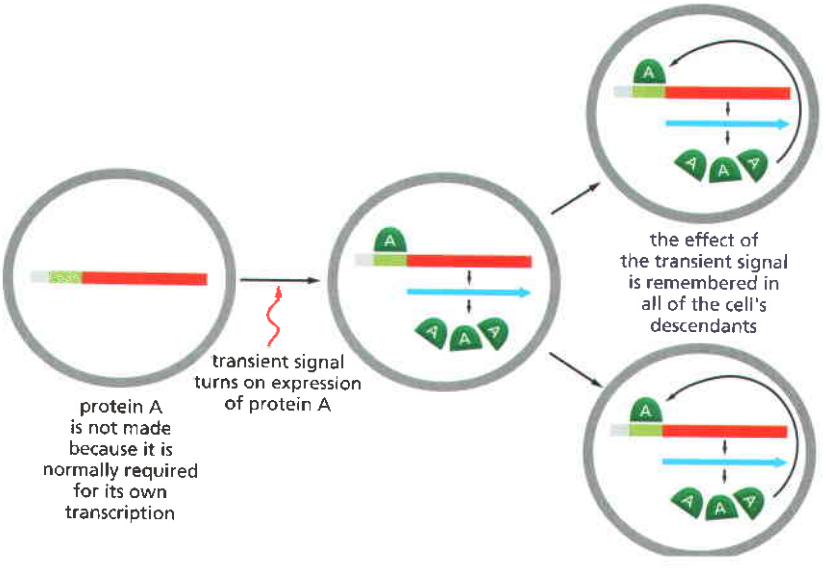


Figure 7–68 Schematic diagram showing how a positive feedback loop can create cell memory. Protein A is a gene regulatory protein that activates its own transcription. All of the descendants of the original cell will therefore "remember" that the progenitor cell had experienced a transient signal that initiated the production of the protein.

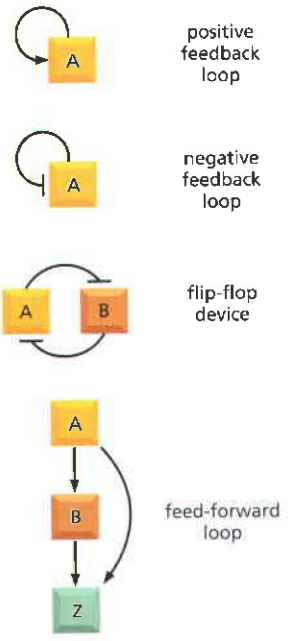


Figure 7–69 Common types of network motifs in transcriptional circuits. A and B represent gene regulatory proteins, arrows indicate positive transcriptional control, and lines with bars depict negative transcriptional control. More detailed descriptions of positive feedback loops and flip-flop devices are given in Figures 7–70 and 7–71, respectively. In feed-forward loops, A and B represent regulatory proteins that both activate the transcription of a target gene, Z.

Transcription Circuits Allow the Cell to Carry Out Logic Operations

Simple gene regulatory switches can be combined to create all sorts of control devices, just as simple electronic switching elements in a computer can be linked to perform many types of operations. The analysis of gene regulatory circuits has revealed that certain simple types of arrangements are found over and over again in cells from widely different species. For example, positive and *negative feedback loops* are especially common in all cells (**Figure 7–69**). As we have seen, the former provides a simple memory device; the latter is often used to keep the expression of a gene close to a standard level regardless of variations in biochemical conditions inside a cell. Suppose, for example, that a transcriptional repressor protein binds to the regulatory region of its own gene and exerts a strong negative feedback, such that transcription occurs at a very low rate when the concentration of repressor protein is above some critical value (determined by its affinity for its DNA binding site), and at a very high rate when it is below the critical value. The concentration of the protein will then be held close to the critical value, since any circumstance that causes a fall below that value will lead to a steep increase in synthesis, and any rise above that value will cause synthesis to switch off. Such adjustments will, however, take time, so that an abrupt change of conditions will cause a disturbance of gene expression that is strong but transient. As we discuss in Chapter 15, the negative feedback system can thus function as a detector of sudden change. Alternatively, if there is a delay in the feedback loop, the result may be spontaneous oscillations in the expression of the gene (see Figure 15–28). The quantitative details of the negative feedback loop determine which of these possible behaviors will occur.

With two or more genes, the possible range of control circuits and circuit behaviors rapidly becomes more complex. Bacteriophage lambda, as we have seen, exemplifies a common type of two-gene circuit that can flip-flop between expression of one gene and expression of the other. Another common circuit arrangement is called a *feed-forward* loop (see Figure 7–69); among other things, this can serve as a filter, responding to input signals that are prolonged but disregarding those that are brief (**Figure 7–70**). A cell can use these various network motifs as miniature logic devices to process information in surprisingly sophisticated ways.

The simple types of devices just illustrated are combined in a typical eucaryotic cell to create exceedingly complex circuits (**Figure 7–71**). Each cell in a developing multicellular organism is equipped with this control machinery, and must, in effect, use the intricate system of interlocking transcriptional switches to compute how it should behave at each time point in response to the many different past and present inputs received. We are only beginning to understand

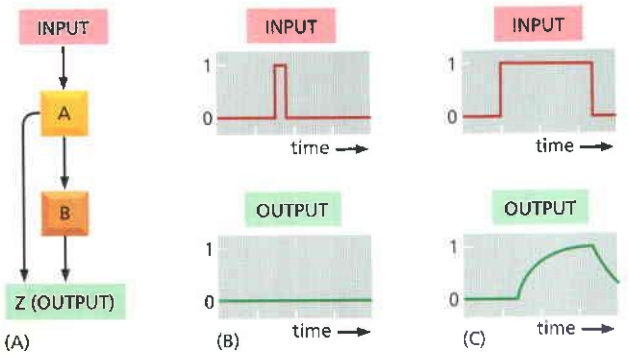


Figure 7–70 How a feed-forward loop can measure the duration of a signal. (A) In this theoretical example the gene activator proteins A and B are both required for transcription of Z, and A becomes active only when an input signal is present. (B) If the input signal to A is brief, A does not stay active long enough for B to accumulate, and the Z gene is not transcribed. (C) If the signal to A persists, B accumulates, A remains active, and Z is transcribed. This arrangement allows the cell to ignore rapid fluctuations of the input signal and respond only to persistent levels. This strategy could be used, for example, to distinguish between random noise and a true signal.

The behavior shown here was computed for one particular set of parameter values describing the quantitative properties of A, B, and Z and their syntheses. With different values of these parameters, feed-forward loops can in principle perform other types of "calculations." Many feed-forward loops have been discovered in cells, and theoretical analysis helps researchers to appreciate and subsequently test the different ways in which they may function. (Adapted from S.S. Shen-Orr, R. Milo, S. Mangan and U. Alon, *Nat. Genet.* 31:64–68, 2002. With permission from Macmillan Publishers Ltd.)

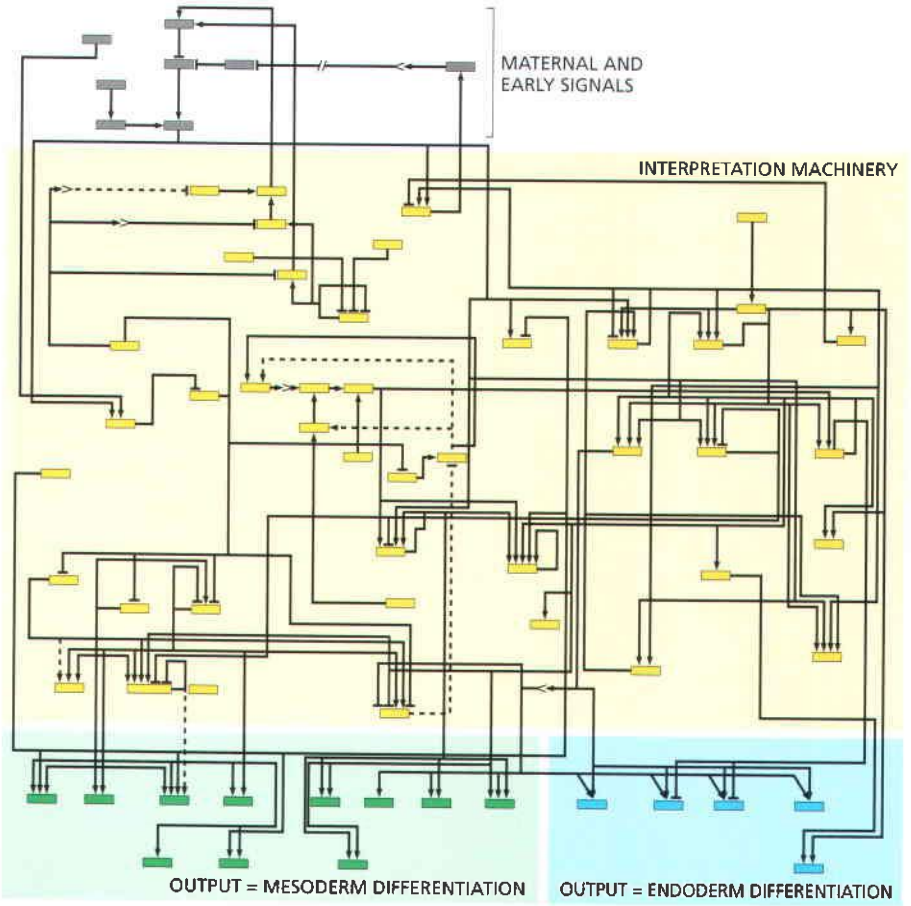


Figure 7–71 The exceedingly complex gene circuit that specifies a portion of the developing sea urchin embryo. Each *colored small box* represents a different gene. Those in *yellow* code for gene regulatory proteins and those in *green* and *blue* code for proteins that give cells of the mesoderm and endoderm, respectively, their specialized characteristics. Genes depicted in *gray* are largely active in the mother and provide the egg with cues needed for proper development. *Arrows* depict instances in which a gene regulatory protein activates the transcription of another gene. Lines ending in *bars* indicate examples of gene repression.

how to study such complex intracellular control networks. Indeed, without quantitative information far more precise and complete than we yet have, it is impossible to predict the behavior of a system such as that shown in Figure 7–71: the circuit diagram by itself is not enough.

Synthetic Biology Creates New Devices from Existing Biological Parts

Our discussion has focused on naturally occurring transcriptional circuits, but it is also instructive to design and construct artificial circuits in the laboratory and introduce them into cells to examine their behavior. **Figure 7–72** shows, for example, how an engineered bacterial cell can switch between three states in a prescribed order, thus functioning as an oscillator or simple clock. The construction of such new devices from existing parts is often termed *synthetic biology*. Scientists use this approach to test whether they truly understand the properties of the component parts; if so, they should be able to combine these parts in novel ways and accurately predict the characteristics of the new device. The fact that these predictions usually fail illustrates how far we are from truly understanding the detailed workings of the cell. There are many large gaps in our knowledge that will require skillful application of the quantitative approaches of the physical sciences to complex biological systems.

Circadian Clocks Are Based on Feedback Loops in Gene Regulation

Life on Earth evolved in the presence of a daily cycle of day and night, and many present-day organisms (ranging from archaea to plants to humans) have come to possess an internal rhythm that dictates different behaviors at different times of day. These behaviors range from the cyclical change in metabolic enzyme

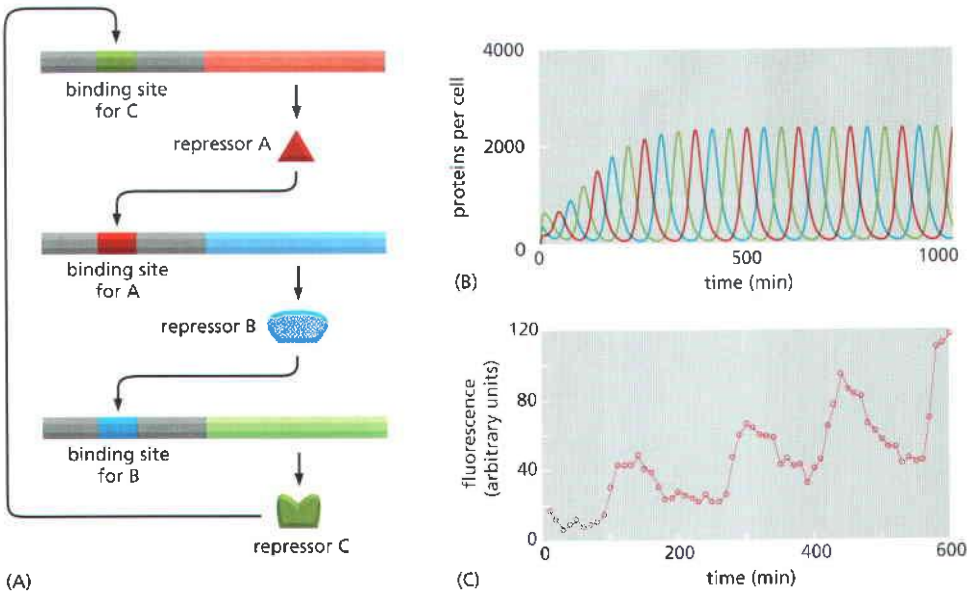


Figure 7–72 A simple gene oscillator or "clock" designed in the laboratory. (A) Recombinant DNA techniques were used to make three artificial genes, each coding for a different bacterial repressor protein, and each controlled by the product of another gene in the set, so as to create a regulatory circuit as shown. These repressors (denoted A, B, and C in the figure) are the Lac repressor (see Figure 7–39), the Tet repressor, which regulates genes in response to tetracycline, and the Lambda repressor (see Figure 7–67). When introduced into a bacterial cell, the three genes form a delayed negative feedback circuit: the product of gene A, for example, acts via genes B and C to indirectly inhibit its own expression. The delayed negative feedback gives rise to oscillations. (B) Computer prediction of the oscillatory behavior. The cell cycles repetitively through a series of states, expressing A, then B, then C, then A again, and so on, as each gene product in turn escapes from inhibition by the previous one and represses the next. (C) Actual oscillations observed in a cell containing the three artificial genes in (A), demonstrated with a fluorescent reporter of the expression of one of these genes. The increasing amplitude of the fluorescence signal reflects the growth of the bacterial cell. (Adapted from M.B. Elowitz and S. Leibler, *Nature* 403:335–338, 2000. With permission from Macmillan Publishers Ltd.)

activities of a fungus to the elaborate sleep–wake cycles of humans. The internal oscillators that control such diurnal rhythms are called circadian clocks.

By carrying its own circadian clock, an organism can anticipate the regular daily changes in its environment and take appropriate action in advance. Of course, the internal clock cannot be perfectly accurate, and so it must be capable of being reset by external cues such as the light of day. Thus, circadian clocks keep running even when the environmental cues (changes in light and dark) are removed, but the period of this free-running rhythm is generally a little less or a little more than 24 hours. External signals indicating the time of day cause small adjustments in the running of the clock, so as to keep the organism in synchrony with its environment. Following more drastic shifts, circadian cycles become gradually reset (entrained) by the new cycle of light and dark, as anyone who has experienced jet lag can attest.

We might expect that the circadian clock in a creature such as a human would itself be a complex multicellular device, with different groups of cells responsible for different parts of the oscillation mechanism. Remarkably, however, it turns out that in almost all organisms, including humans, the timekeepers are individual cells. Thus, a clock that operates in each member of a specialized group of brain cells (the SCN cells in the suprachiasmatic nucleus of the hypothalamus) controls our diurnal cycles of sleeping and waking, body temperature, and hormone release. Even if these cells are removed from the brain and dispersed in a culture dish, they will continue to oscillate individually, showing a cyclic pattern of gene expression with a period of approximately 24 hours. In the intact body, the SCN cells receive neural cues from the retina, entraining them to the daily cycle of light and dark, and they send information about the time of day to another brain area, the pineal gland, which relays the time signal to the rest of the body by releasing the hormone melatonin in time with the clock.

Although the SCN cells have a central role as timekeepers in mammals, they are not the only cells in the mammalian body that have an internal circadian rhythm or an ability to reset it in response to light. Similarly, in *Drosophila*, many different types of cells, including those of the thorax, abdomen, antenna, leg, wing, and testis all continue a circadian cycle when they have been dissected away from the rest of the fly. The clocks in these isolated tissues, like those in the SCN cells, can be reset by externally imposed light and dark cycles.

The working of circadian clocks, therefore, is a fundamental problem in cell biology. Although we do not yet understand all the details, studies in a wide variety of organisms have revealed many of the basic principles and molecular components. For animals, much of what we know has come from searches in *Drosophila* for mutations that make the fly's circadian clock run fast, or slow, or not at all; and this work has led to the discovery that many of the same components are involved in the circadian clock of mammals.

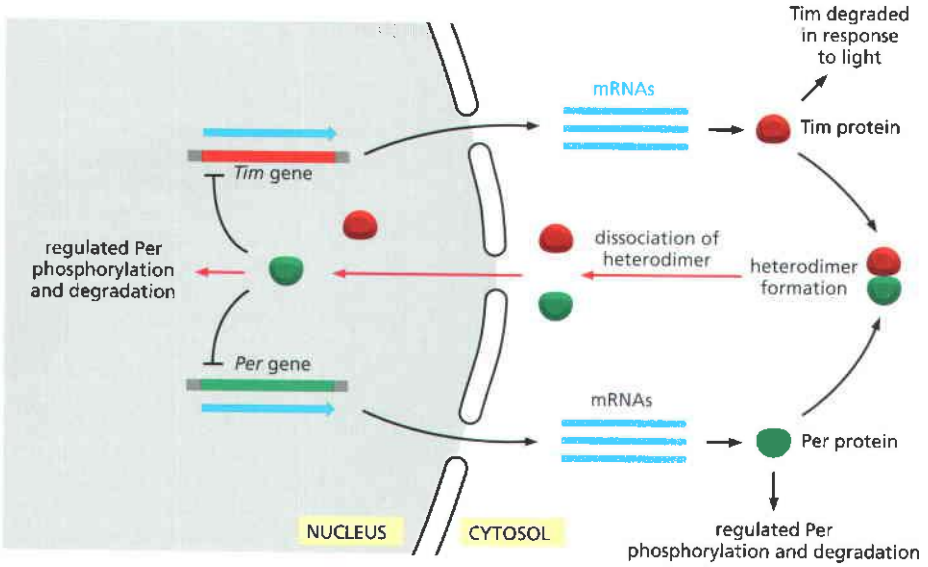


Figure 7–73 Simplified outline of the mechanism of the circadian clock in *Drosophila* cells. A central feature of the clock is the periodic accumulation and decay of two gene regulatory proteins, Tim (short for timeless, based on the phenotype of a gene mutation) and Per (short for period). The mRNAs encoding these proteins are translated in the cytosol, and, when each protein has accumulated to critical levels, they form a heterodimer. After a time delay, the heterodimer dissociates and Tim and Per are transported into the nucleus, where they regulate a number of gene products that mediate effects of the clock. Once in the nucleus, Per also represses the *Tim* and *Per* genes, creating a feedback system that causes the levels of Tim and Per to fall. In addition to this transcriptional feedback, the clock depends on a set of other proteins. For example, the controlled degradation of Per indicated in the diagram imposes delays in the periodic accumulation of Tim and Per, which are crucial to the functioning of the clock. Steps at which specific delays are imposed are shown in *red*.

Entrainment (or resetting) of the clock occurs in response to new light–dark cycles. Although most *Drosophila* cells do not have true photoreceptors, light is sensed by intracellular flavoproteins, also called cryptochromes. In the presence of light, these proteins associate with the Tim protein and cause its degradation, thereby resetting the clock. (Adapted from J.C. Dunlap, *Science* 311:184–186, 2006. With permission from AAAS.)

The mechanism of the clock in *Drosophila* is briefly outlined in **Figure 7–73**. At the heart of the oscillator is a transcriptional feedback loop that has a time delay built into it: accumulation of certain key gene products switches off the transcription of their genes, but with a delay, so that—crudely speaking—the cell oscillates between a state in which the products are present and transcription is switched off, and one in which the products are absent and transcription is switched on.

Despite the relative simplicity of the basic principle behind circadian clocks, the details are complex. One reason for this complexity is that clocks must be buffered against changes in temperature, which typically speed up or slow down macromolecular association. They must also run accurately but be capable of being reset. Although it is not yet understood how biological clocks run at a constant speed despite changes in temperature, the mechanism for resetting the *Drosophila* clock is the light-induced destruction of one of the key gene regulatory proteins (see Figure 7–73).

A Single Gene Regulatory Protein Can Coordinate the Expression of a Set of Genes

Cells need to be able to switch genes on and off individually but they also need to coordinate the expression of large groups of different genes. For example, when a quiescent eucaryotic cell receives a signal to divide, many hitherto unexpressed genes are turned on together to set in motion the events that lead eventually to cell division (discussed in Chapter 17). One way bacteria coordinate the expression of a set of genes is to cluster them together in an *operon* under control of a single promoter (see Figure 7–34). In eucaryotes, however, each gene is transcribed from a separate promoter.

How, then, do eucaryotes coordinate gene expression? This is an especially important question because, as we have seen, most eucaryotic gene regulatory proteins act as part of a regulatory protein committee, all of whose members are necessary to express the gene in the right cell, at the right time, in response to the proper signals, and to the proper level. How, then, can a eucaryotic cell rapidly and decisively switch whole groups of genes on or off?

The answer is that even though control of gene expression is combinatorial, the effect of a single gene regulatory protein can still be decisive in switching any particular gene on or off, simply by completing the combination needed to maximally activate or repress that gene. This situation is analogous to dialing in the final number of a combination lock: the lock will spring open with only this simple addition if all of the other numbers have been previously entered. Moreover,

the same number can complete the combination for many different locks. Analogously, the addition of a particular protein can turn on many different genes.

An example in humans is the control of gene expression by the *glucocorticoid receptor protein.* To bind to regulatory sites in DNA, this gene regulatory protein must first form a complex with a molecule of a glucocorticoid steroid hormone, such as cortisol (see Figure 15–13). The body releases this hormone during times of starvation and intense physical activity, and among its other activities, it stimulates liver cells to increase the production of glucose from amino acids and other small molecules. To respond in this way, liver cells increase the expression of many different genes that code for metabolic enzymes and other products. Although these genes all have different and complex control regions, their maximal expression depends on the binding of the hormone–glucocorticoid receptor complex to a regulatory site in the DNA of each gene. When the body has recovered and the hormone is no longer present, the expression of each of these genes drops to its normal level in the liver. In this way a single gene regulatory protein can control the expression of many different genes (**Figure 7–74**).

The effects of the glucocorticoid receptor are not confined to cells of the liver. In other cell types, activation of this gene regulatory protein by hormone also causes changes in the expression levels of many genes; the genes affected, however, are often different from those affected in liver cells. As we have seen, each cell type has an individualized set of gene regulatory proteins, and because of combinatorial control, these critically influence the action of the glucocorticoid receptor. Because the receptor is able to assemble with many different sets of cell-type-specific gene regulatory proteins, switching it on with hormone produces a different spectrum of effects in each cell type.

Expression of a Critical Gene Regulatory Protein Can Trigger the Expression of a Whole Battery of Downstream Genes

The ability to switch many genes on or off coordinately is important not only in the day-to-day regulation of cell function. It is also the means by which eucaryotic cells differentiate into specialized cell types during embryonic development. The development of muscle cells provides a striking example.

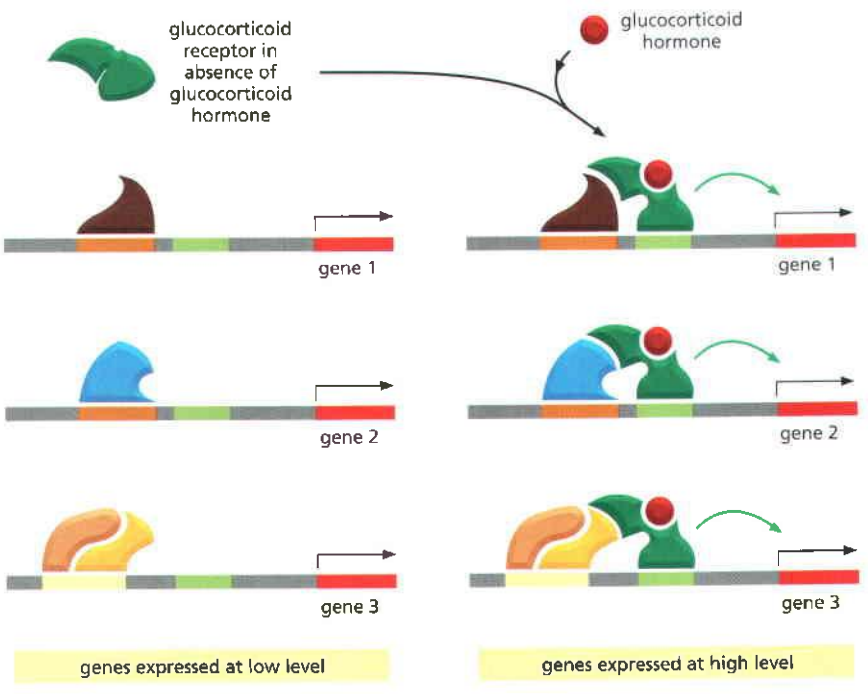


Figure 7–74 A single gene regulatory protein can coordinate the expression of several different genes. The action of the glucocorticoid receptor is illustrated schematically. On the *left* is a series of genes, each of which has various gene activator proteins bound to its regulatory region. However, these bound proteins are not sufficient on their own to fully activate transcription. On the *right* is shown the effect of adding an additional gene regulatory protein—the glucocorticoid receptor in a complex with glucocorticoid hormone—that can bind to the regulatory region of each gene. The glucocorticoid receptor completes the combination of gene regulatory proteins required for maximal initiation of transcription, and the genes are now switched on as a set. In the absence of the hormone, the glucocorticoid receptor is unavailable to bind to DNA.

In addition to activating gene expression, the hormone-bound form of the glucocorticoid receptor represses transcription of certain genes, depending on the gene regulatory proteins already present on their control regions. The effect of the glucocorticoid receptor on any given gene therefore depends upon the type of cell, the gene regulatory proteins contained within it, and the regulatory region of the gene. The structure of the DNA-binding portion of the glucocorticoid receptor is shown in Figure 7–16.

As described in Chapter 16, a mammalian skeletal muscle cell is a highly distinctive giant cell, formed by the fusion of many muscle precursor cells called *myoblasts*, and therefore containing many nuclei. The mature muscle cell synthesizes a large number of characteristic proteins, including specific types of actin, myosin, tropomyosin, and troponin (all part of the contractile apparatus), creatine phosphokinase (for the specialized metabolism of muscle cells), and acetylcholine receptors (to make the membrane sensitive to nerve stimulation). In proliferating myoblasts, these muscle-specific proteins and their mRNAs are absent or are present in very low concentrations. As myoblasts begin to fuse with one another, the corresponding genes are all switched on coordinately as part of a general transformation of the pattern of gene expression.

This entire program of muscle differentiation can be triggered in cultured skin fibroblasts and certain other cell types by introducing any one of a family of helix–loop–helix proteins—the so-called myogenic proteins (MyoD, Myf5, MyoG, and Mrf4)—that are normally expressed only in muscle cells (**Figure 7–75**A). Binding sites for these regulatory proteins are present in the regulatory DNA sequences adjacent to many muscle-specific genes, and the myogenic proteins thereby directly activate the transcription of these genes. In addition, the myogenic proteins stimulate their own transcription as well as that of various other gene regulatory proteins involved in muscle development, creating an elaborate series of positive feedback and feed-forward loops that amplify and maintain the muscle developmental program, even after the initiating signal has disappeared (Figure 7–75B; see also Chapter 22).

It is probable that those cell types that are converted to muscle cells by the addition of myogenic proteins have already accumulated a number of gene regulatory proteins that can cooperate with the myogenic proteins to switch on muscle-specific genes. Other cell types fail to be converted to muscle by myogenin or its relatives; these cells presumably have not accumulated the other gene regulatory proteins required.

The conversion of one cell type (fibroblast) to another (skeletal muscle) by a single gene regulatory protein reemphasizes one of the most important principles discussed in this chapter: differences in gene expression can produce dramatic differences between cell types—in size, shape, chemistry, and function.

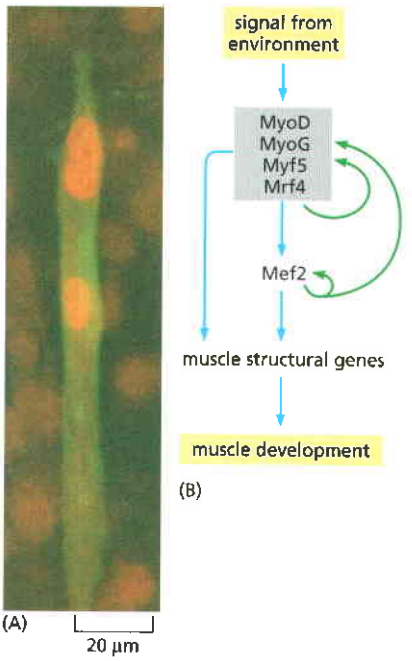


Figure 7–75 Role of the myogenic regulatory proteins in muscle development. (A) The effect of expressing the MyoD protein in fibroblasts. As shown in this immunofluorescence micrograph, fibroblasts from the skin of a chick embryo have been converted to muscle cells by the experimentally induced expression of the *MyoD* gene. The fibroblasts that have been induced to express the *MyoD* gene have fused to form elongated multinucleate muscle-like cells, which are stained *green* with an antibody that detects a muscle-specific protein. Fibroblasts that do not express the *MyoD* gene are barely visible in the background. (B) Simplified scheme showing some of the gene regulatory proteins involved in skeletal muscle development. External signals result in the synthesis of the four closely related myogenic gene regulatory proteins, MyoD, Myf5, MyoG, and Mrf4. These gene regulatory proteins activate their own as well as each other's synthesis in a complex series of feedback loops, only some of which are shown in the figure. These proteins in turn directly activate transcription of muscle structural genes as well as the *Mef2* gene, which encodes an additional gene regulatory protein. Mef2 acts in combination with the myogenic proteins in a feed-forward loop to further activate the transcription of muscle structural genes, as well as forming an additional positive feedback loop that helps to maintain transcription of the myogenic genes. (A, courtesy of Stephen Tapscott and Harold Weintraub; B, adapted from J.D. Molkentin and E.N. Olson, *Proc. Natl Acad. Sci. U.S.A.* 93:9366–9373, 1996. With permission from National Academy of Sciences.)

Combinatorial Gene Control Creates Many Different Cell Types in Eucaryotes

We have already discussed how multiple gene regulatory proteins can act in combination to regulate the expression of an individual gene. But, as the example of the myogenic proteins shows, combinatorial gene control means more than this: not only does each gene respond to many gene regulatory proteins that control it, but each regulatory protein contributes to the control of many genes. Moreover, although some gene regulatory proteins are specific to a single cell type, most are switched on in a variety of cell types, at several sites in the body, and at several times in development. This point is illustrated schematically in **Figure 7–76**, which shows how combinatorial gene control makes it possible to generate a great deal of biological complexity even with relatively few gene regulatory proteins.

With combinatorial control, a given gene regulatory protein does not necessarily have a single, simply definable function as commander of a particular battery of genes or specifier of a particular cell type. Rather, gene regulatory proteins can be likened to the words of a language: they are used with different meanings in a variety of contexts and rarely alone; it is the well-chosen combination that conveys the information that specifies a gene regulatory event.

One requirement of combinatorial control is that many gene regulatory proteins must be able to work together to influence the final rate of transcription. Experiments demonstrate that even unrelated gene regulatory proteins from widely different eucaryotic species can cooperate when introduced into the same cell. This situation reflects the high degree of conservation of the transcription

machinery. It seems that the multifunctional, combinatorial mode of action of gene regulatory proteins has put a tight constraint on their evolution: they must interlock with other gene regulatory proteins, the general transcription factors, Mediator, RNA polymerase, and the chromatin-modifying enzymes.

An important consequence of combinatorial gene control is that the effect of adding a new gene regulatory protein to a cell will depend on the cell's past history, since this history will determine which gene regulatory proteins are already present. Thus, during development a cell can accumulate a series of gene regulatory proteins that need not initially alter gene expression. The addition of the final members of the requisite combination of gene regulatory proteins completes the regulatory message, and can lead to large changes in gene expression. Such a scheme, as we have seen, helps to explain how the addition of a single regulatory protein to a fibroblast can produce the dramatic transformation of the fibroblast into a muscle cell. It also can account for the important difference, discussed in Chapter 22, between the process of *cell determination*—where a cell becomes committed to a particular developmental fate—and the process of *cell differentiation*, in which a committed cell expresses its specialized character.

A Single Gene Regulatory Protein Can Trigger the Formation of an Entire Organ

We have seen that even though combinatorial control is the norm for eucaryotic genes, a single gene regulatory protein, if it completes the appropriate combination, can be decisive in switching a whole set of genes on or off, and we have

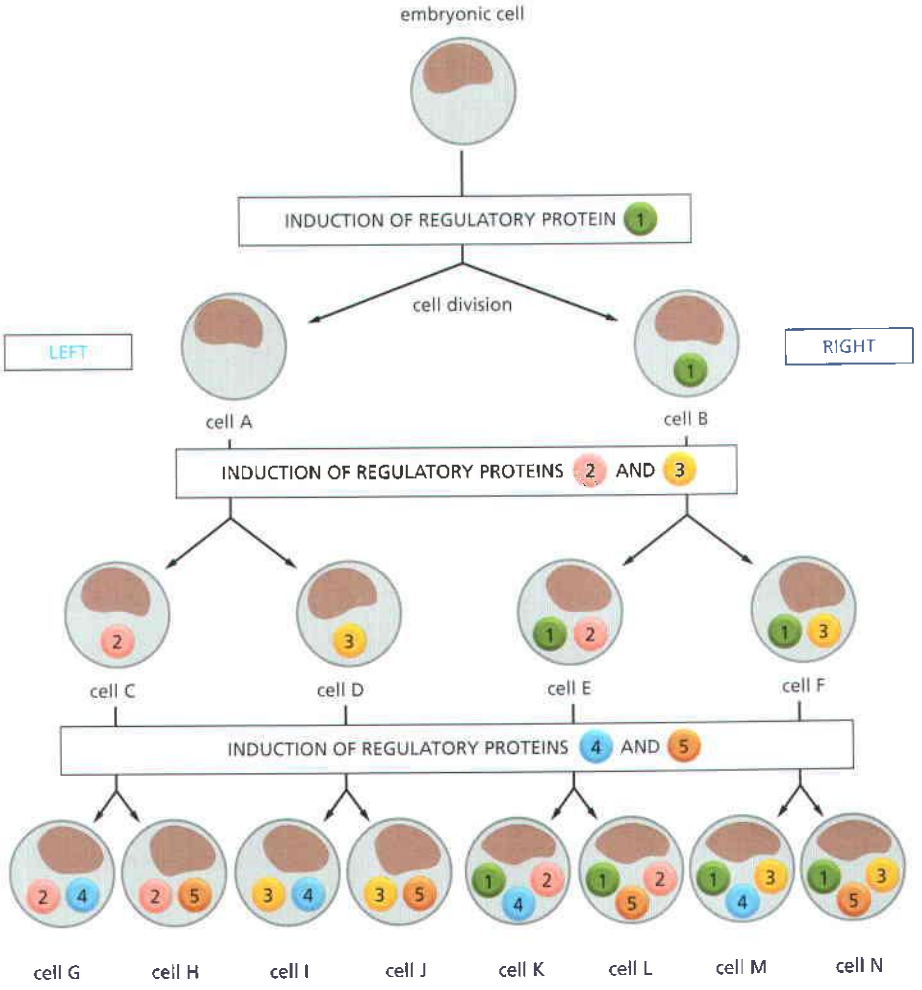


Figure 7–76 The importance of combinatorial gene control for development. Combinations of a few gene regulatory proteins can generate many cell types during development. In this simple, idealized scheme a "decision" to make one of a pair of different gene regulatory proteins (shown as numbered *circles*) is made after each cell division. Sensing its relative position in the embryo, the daughter cell toward the *left side* of the embryo is always induced to synthesize the even-numbered protein of each pair, while the daughter cell toward the *right side* of the embryo is induced to synthesize the odd-numbered protein. The production of each gene regulatory protein is assumed to be self-perpetuating once it has become initiated (see Figure 7–68). In this way, through cell memory, the final combinatorial specification is built up step by step. In this purely hypothetical example, five different gene regulatory proteins have created eight final cell types (G–N).

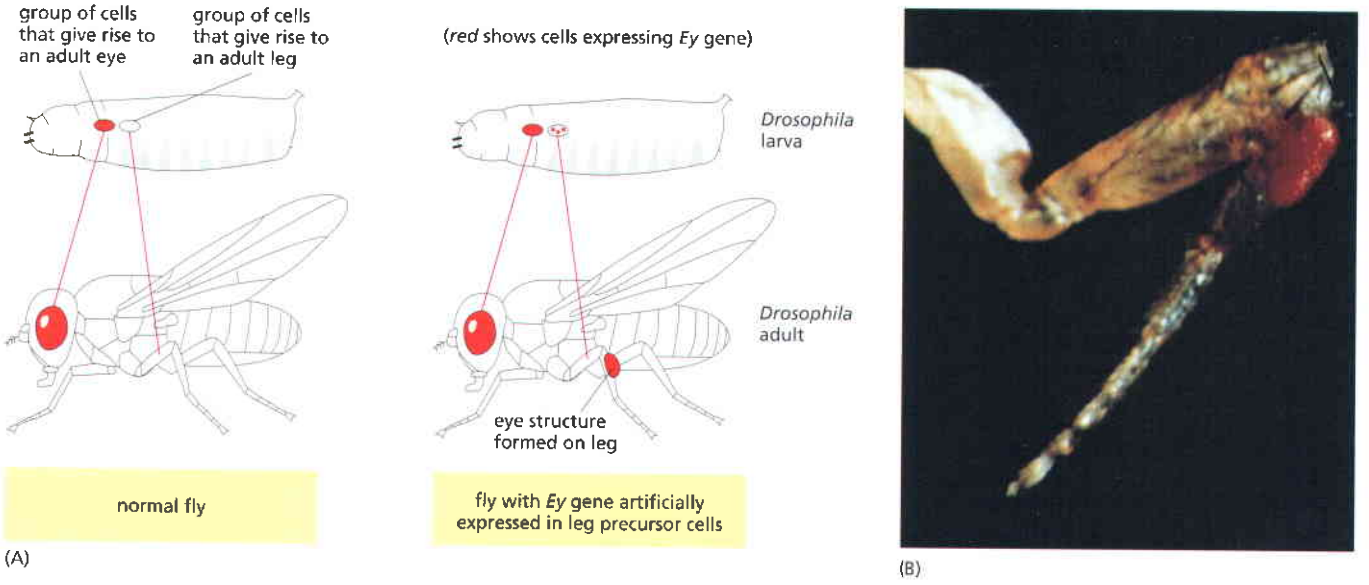


Figure 7–77 Expression of the *Drosophila Ey* gene in precursor cells of the leg triggers the development of an eye on the leg. (A) Simplified diagrams showing the result when a fruit fly larva contains either the normally expressed *Ey* gene *(left)* or an *Ey* gene that is additionally expressed artificially in cells that normally give rise to leg tissue *(right)*. (B) Photograph of an abnormal leg that contains a misplaced eye (see also Figure 22–2). (B, courtesy of Walter Gehring.)

seen how this can convert one cell type into another. A dramatic extension of the principle comes from studies of eye development in *Drosophila*, mice, and humans. Here, a gene regulatory protein, called Ey (short for Eyeless) in flies and Pax6 in vertebrates, is crucial. When expressed in the proper context, Ey can trigger the formation of not just a single cell type but a whole organ (an eye), composed of different types of cells, all properly organized in three-dimensional space.

The most striking evidence for the role of Ey comes from experiments in fruit flies in which the *Ey* gene is artificially expressed early in development in groups of cells that normally will go on to form leg parts. This abnormal gene expression causes eyes to develop in the legs (**Figure 7–77**).

The *Drosophila* eye is composed of thousands of cells, and the question of how a regulatory protein coordinates the construction of a whole organ is a central topic in *developmental biology*. As discussed in Chapter 22, it involves cell–cell interactions as well as intracellular gene regulatory proteins. Here, we note that Ey directly controls the expression of many other genes by binding to their regulatory regions. Some of the genes controlled by Ey themselves code for gene regulatory proteins that, in turn, control the expression of other genes. Moreover, some of these regulatory gene products act back on *Ey* itself to create a positive feedback loop that ensures the continued synthesis of the Ey protein as the cells divide and further differentiate (**Figure 7–78**). In this way, the action of just one regulatory protein can permanently turn on a cascade of gene regulatory proteins and cell–cell interaction mechanisms, whose actions result in an organized group of many different types of cells. One can begin to imagine how, by repeated applications of this principle, a complex organism is assembled piece by piece.

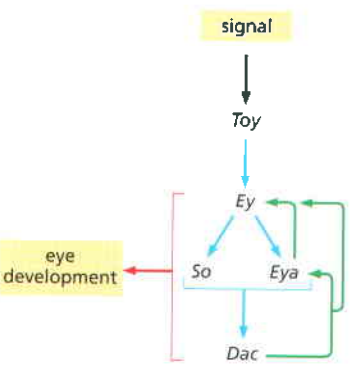


Figure 7–78 Gene regulatory proteins that specify eye development in *Drosophila*. *Toy (Twin of eyeless)* and *Ey (Eyeless)* encode similar gene regulatory proteins, Toy and Ey, either of which, when ectopically expressed, can trigger eye development. In normal eye development, expression of *Ey* requires the *Toy* gene. Once its transcription is activated by Toy, Ey activates the transcription of *So (Sine oculis)* and *Eya (Eyes absent)*, which act together to switch on the *Dac (Dachshund)* gene. As indicated by the *green arrows*, some of the gene regulatory proteins form a series of interlocking positive feedback loops that reinforce the initial commitment to eye development. The Ey protein is known to bind directly to numerous target genes for eye development, including those encoding lens crystallins, rhodopsins, and other photoreceptor proteins. (Adapted from T. Czerny et al., *Mol. Cell* 3:297–307, 1999. With permission from Elsevier.)

The Pattern of DNA Methylation Can Be Inherited When Vertebrate Cells Divide

Thus far, we have emphasized the regulation of gene transcription by proteins that associate with DNA. However, DNA itself can be covalently modified, and in the following sections we shall see that this, too, provides opportunities for the regulation of gene expression. In vertebrate cells, the methylation of cytosine provides a powerful mechanism through which gene expression patterns are passed on to progeny cells. The methylated form of cytosine, 5-methylcytosine (5-methyl C), has the same relation to cytosine that thymine has to uracil, and the modification likewise has no effect on base-pairing (**Figure 7–79**). **DNA methylation** in vertebrate DNA is restricted to cytosine (C) nucleotides in the sequence CG, which is base-paired to exactly the same sequence (in opposite orientation) on the other strand of the DNA helix. Consequently, a simple mechanism permits the existing pattern of DNA methylation to be inherited directly by the daughter DNA strands. An enzyme called *maintenance methyltransferase* acts preferentially on those CG sequences that are base-paired with a CG sequence that is already methylated. As a result, the pattern of DNA methylation on the parental DNA strand serves as a template for the methylation of the daughter DNA strand, causing this pattern to be inherited directly following DNA replication (**Figure 7–80**).

The stable inheritance of DNA methylation patterns can be explained by maintenance DNA methyltransferases. DNA methylation patterns, however, are dynamic during vertebrate development. Shortly after fertilization there is a genome-wide wave of demethylation, when the vast majority of methyl groups are lost from the DNA. This demethylation may occur either by suppression of maintenance DNA methyltransferase activity, resulting in the passive loss of methyl groups during each round of DNA replication, or by a specific demethylating enzyme. Later in development, new methylation patterns are established by several *de novo DNA methyltransferases* that are directed to DNA by sequence-specific DNA-binding proteins where they modify adjacent unmethylated CG nucleotides. Once the new patterns of methylation are established, they can be propagated through rounds of DNA replication by the maintenance methyl transferases.

DNA methylation has several uses in the vertebrate cell. Perhaps its most important role is to work in conjunction with other gene expression control mechanisms to establish a particularly efficient form of gene repression that can be faithfully passed on to progeny cells (**Figure 7–81**). This combination of mechanisms ensures that unneeded eucaryotic genes can be repressed to very high degrees. For example, the rate at which a vertebrate gene is transcribed can vary 10^6-fold between one tissue and another. The unexpressed vertebrate genes are much less "leaky" in terms of transcription than bacterial genes, in which the largest known differences in transcription rates between expressed and unexpressed gene states are about 1000-fold.

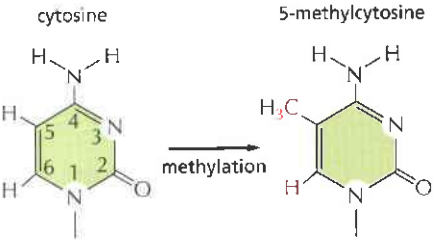


Figure 7–79 Formation of 5-methylcytosine occurs by methylation of a cytosine base in the DNA double helix. In vertebrates this event is confined to selected cytosine (C) nucleotides located in the sequence CG.

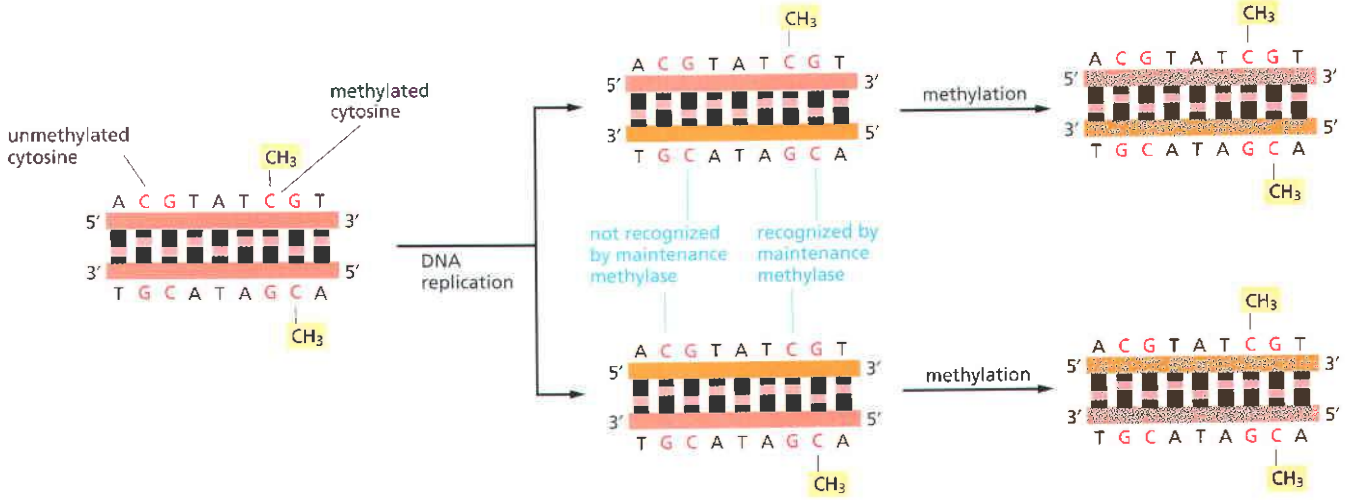


Figure 7–80 How DNA methylation patterns are faithfully inherited. In vertebrate DNAs a large fraction of the cytosine nucleotides in the sequence CG are methylated (see Figure 7–79). Because of the existence of a methyl-directed methylating enzyme (the maintenance methyltransferase), once a pattern of DNA methylation is established, that pattern of methylation is inherited in the progeny DNA, as shown.

How DNA methylation helps to repress gene expression is not understood in detail, but two general mechanisms have emerged. DNA methylation of the promoter region of a gene or of its regulatory sequences can interfere directly with the binding of proteins required for transcription initiation. In addition, the cell has a repertoire of proteins that specifically bind to methylated DNA (see Figure 7–81), thereby blocking access of other proteins. One reflection of the importance of DNA methylation to humans is the widespread involvement of errors in this mechanism in cancer progression (see Chapter 20).

We shall return to the topic of gene silencing by DNA methylation later in this chapter, when we discuss X-chromosome inactivation and other examples of large-scale gene silencing. First, however, we describe some of the other ways in which DNA methylation affects our genomes.

Genomic Imprinting Is Based on DNA Methylation

Mammalian cells are diploid, containing one set of genes inherited from the father and one set from the mother. The expression of a small minority of genes depends on whether they have been inherited from the mother or the father: while the paternally inherited gene copy is active, the maternally inherited gene copy is silent, or vice-versa. This phenomenon is called **genomic imprinting**. The gene for *insulin-like growth factor-2* (*Igf2*) is a well-studied example of an

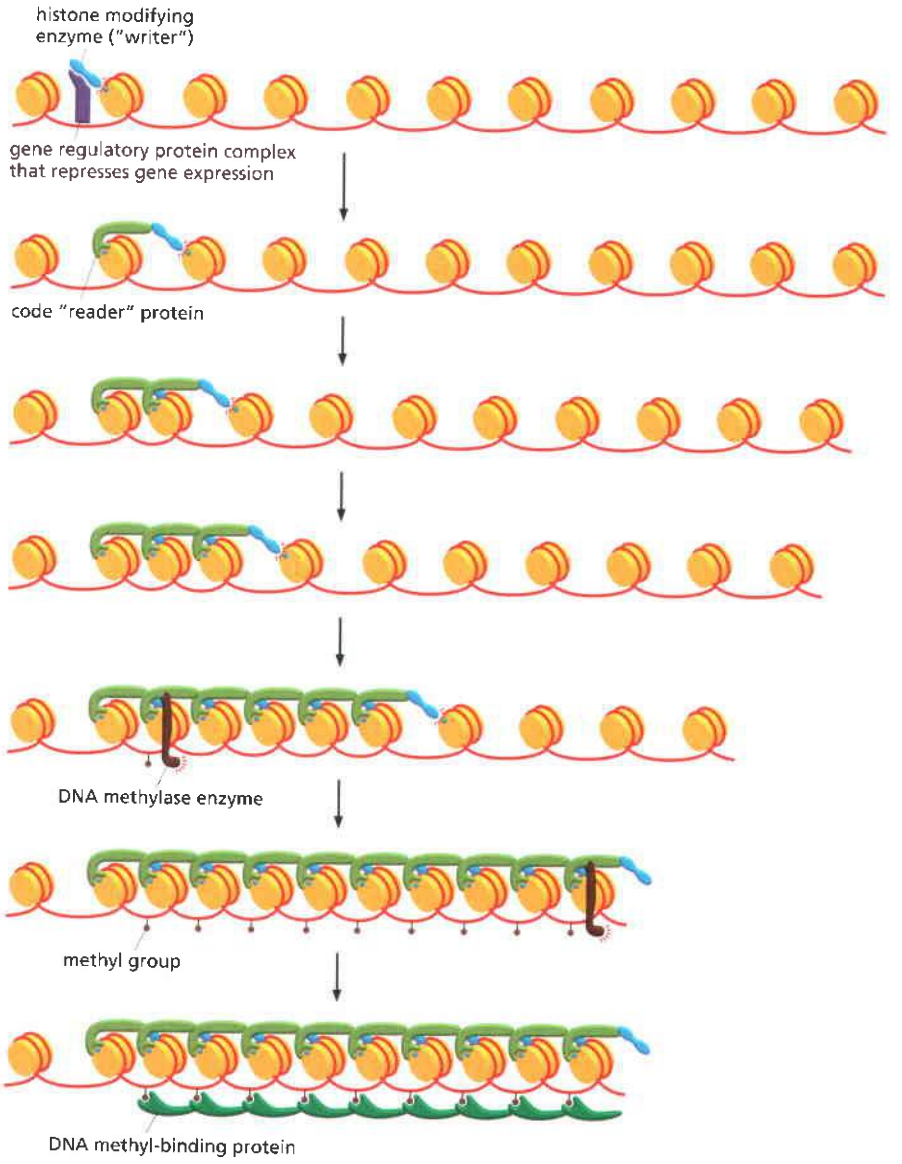


Figure 7–81 Multiple mechanisms contribute to stable gene repression. In this schematic example, histone reader and writer proteins, under the direction of gene regulatory proteins, establish a repressive form of chromatin. A *de novo* DNA methylase is attracted by the histone reader and methylates nearby cytosines in DNA, which are, in turn, bound by DNA methyl-binding proteins. During DNA replication, some of the modified (*blue* dot) histones will be inherited by one daughter chromosome, some by the other, and in each daughter they can induce reconstruction of the same pattern of chromatin modifications (see Figure 5–39). At the same time, the mechanism shown in Figure 7–80 will cause both daughter chromosomes to inherit the same methylation pattern. The two inheritance mechanisms will be mutually reinforcing, if DNA methylation stimulates the activity of the histone writer. This scheme can account for the inheritance by daughter cells of both the histone and the DNA modifications. It can also explain the tendency of some chromatin modifications to spread along a chromosome (see Figure 4–45).

imprinted gene. *Igf2* is required for prenatal growth, and mice that do not express *Igf2* at all are born half the size of normal mice. However, only the paternal copy of *Igf2* is transcribed, and only this gene copy matters for the phenotype. As a result, mice with a mutated paternally derived *Igf2* gene are stunted, while mice with a mutated maternally derived *Igf2* gene are normal.

In the early embryo, genes subject to imprinting are marked by methylation according to whether they were derived from a sperm or an egg chromosome. In this way, DNA methylation is used as a mark to distinguish two copies of a gene that may be otherwise identical (**Figure 7–82**). Because imprinted genes are somehow protected from the wave of demethylation that takes place shortly after fertilization (see p. 467), this mark enables somatic cells to "remember" the

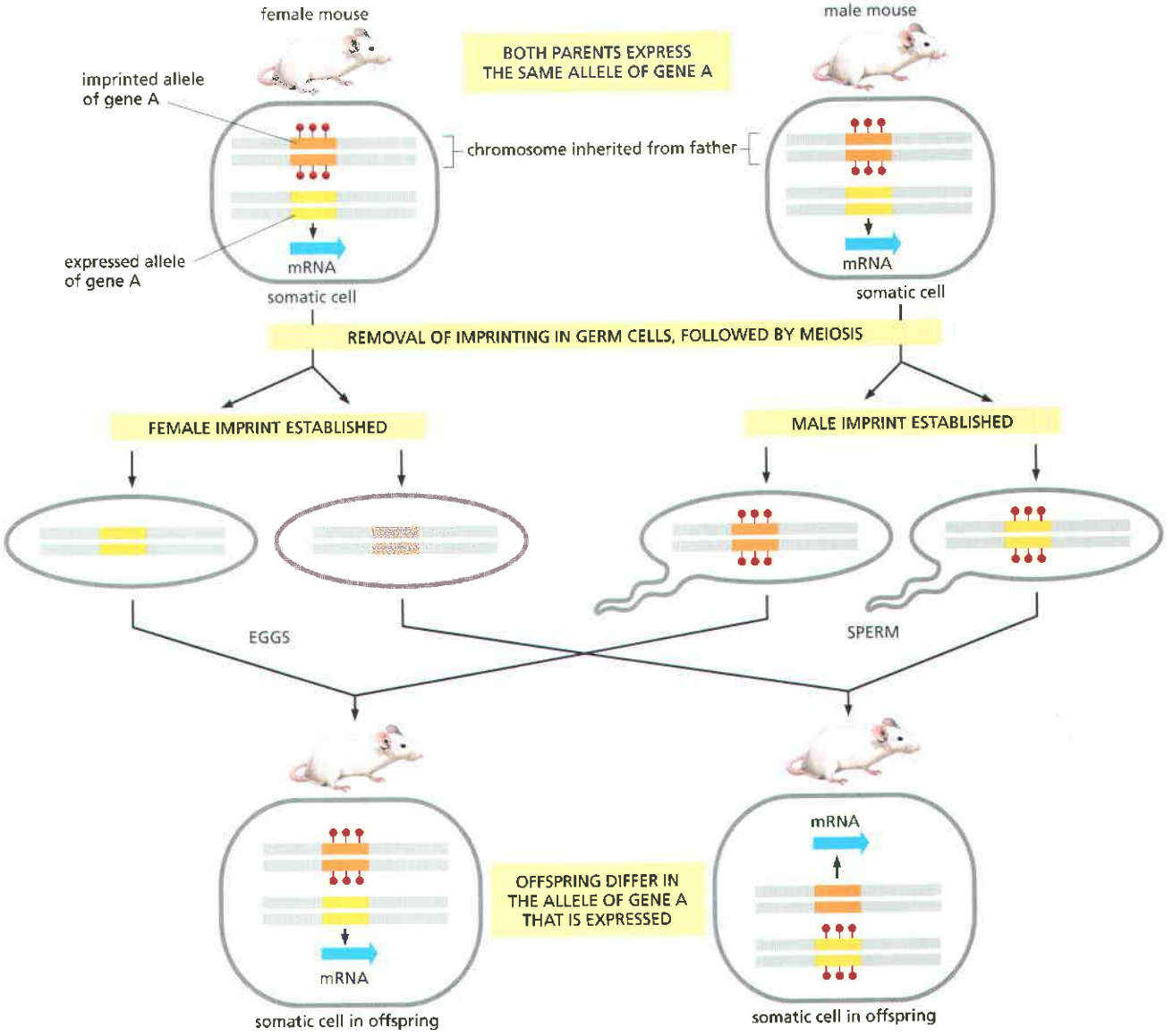


Figure 7–82 Imprinting in the mouse. The *top* portion of the figure shows a pair of homologous chromosomes in the somatic cells of two adult mice, one male and one female. In this example, both mice have inherited the top homolog from their father and the bottom homolog from their mother, and the paternal copy of a gene subject to imprinting (indicated in *orange*) is methylated, preventing its expression. The maternally derived copy of the same gene (*yellow*) is expressed. The remainder of the figure shows the outcome of a cross between these two mice. During germ cell formation, but before meiosis, the imprints are erased and then, much later in germ cell development, they are reimposed in a sex-specific pattern (*middle* portion of figure). In eggs produced from the female, neither allele of the A gene is methylated. In sperm from the male, both alleles of gene A are methylated. Shown at the *bottom* of the figure are two of the possible imprinting patterns inherited by the progeny mice; the mouse on the *left* has the same imprinting pattern as each of the parents, whereas the mouse on the *right* has the opposite pattern. If the two alleles of A gene are distinct, these different imprinting patterns can cause phenotypic differences in the progeny mice, even though they carry exactly the same DNA sequences of the two A gene alleles. Imprinting provides an important exception to classical genetic behavior, and several hundred mouse genes are thought to be affected in this way. However, the majority of mouse genes are not imprinted, and therefore the rules of Mendelian inheritance apply to most of the mouse genome.

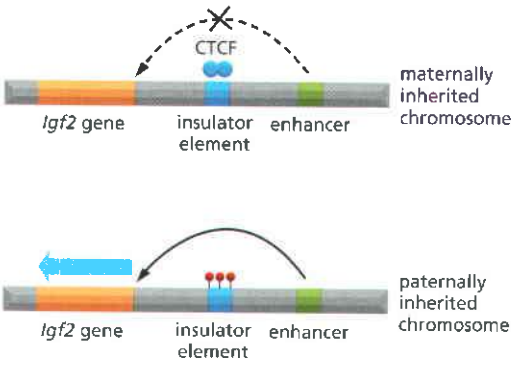


Figure 7–83 Mechanism of imprinting of the mouse *Igf2* gene. On chromosomes inherited from the female, a protein called CTCF binds to an insulator (see Figure 7–62), blocking communication between the enhancer *(green)* and the *Igf2* gene *(orange)*. IGF2 is therefore not expressed from the maternally inherited chromosome. Because of imprinting, the insulator on the male-derived chromosome is methylated; this inactivates the insulator, by blocking the binding of the CTCF protein, and allows the enhancer to activate transcription of the *Igf2* gene. In other examples of imprinting, methylation blocks gene expression by interfering with the binding of proteins required for a gene's transcription.

parental origin of each of the two copies of the gene and to regulate their expression accordingly. In most cases, the methyl imprint silences nearby gene expression. In some cases, however, the methyl imprint can activate expression of a gene. In the case of *Igf2*, for example, methylation of an insulator element (see Figure 7–62) on the paternally derived chromosome blocks its function and allows a distant enhancer to activate transcription of the *Igf2* gene. On the maternally derived chromosome, the insulator is not methylated and the *Igf2* gene is therefore not transcribed (**Figure 7–83**).

Why imprinting should exist at all is a mystery. In vertebrates, it is restricted to placental mammals, and many of the imprinted genes are involved in fetal development. One idea is that imprinting reflects a middle ground in the evolutionary struggle between males to produce larger offspring and females to limit offspring size. Whatever its purpose might be, imprinting provides startling evidence that features of DNA other than its sequence of nucleotides can be inherited.

CG-Rich Islands Are Associated with Many Genes in Mammals

Because of the way in which DNA repair enzymes work, methylated C nucleotides in the genome tend to be eliminated in the course of evolution. Accidental deamination of an unmethylated C gives rise to U (see Figure 5–45), which is not normally present in DNA and thus is recognized easily by the DNA repair enzyme uracil DNA glycosylase, excised, and then replaced with a C (as discussed in Chapter 5). But accidental deamination of a 5-methyl C cannot be repaired in this way, for the deamination product is a T and so is indistinguishable from the other, nonmutant T nucleotides in the DNA. Although a special repair system exists to remove these mutant T nucleotides, many of the deaminations escape detection, so that those C nucleotides in the genome that are methylated tend to mutate to T over evolutionary time.

During the course of evolution, more than three out of every four CGs have been lost in this way, leaving vertebrates with a remarkable deficiency of this dinucleotide. The CG sequences that remain are very unevenly distributed in the genome; they are present at 10–20 times their average density in selected regions, called **CG islands**, which are 1000–2000 nucleotide pairs long. These islands, with some important exceptions, seem to remain unmethylated in all cell types. They often surround the promoters of the so-called *housekeeping genes*—those genes that code for the many proteins that are essential for cell viability and are therefore expressed in most cells (**Figure 7–84**).

The distribution of CG islands (also called CpG islands to distinguish the CG dinucleotides from the CG base pair) can be explained if we assume that CG methylation was adopted in vertebrates primarily as a way of maintaining DNA in a transcriptionally inactive state (see Figure 7–81). In vertebrates, new methyl-C to T mutations can be transmitted to the next generation only if they occur in the germ line, the cell lineage that gives rise to sperm or eggs. Most of the DNA in vertebrate germ cells is inactive and highly methylated. Over long

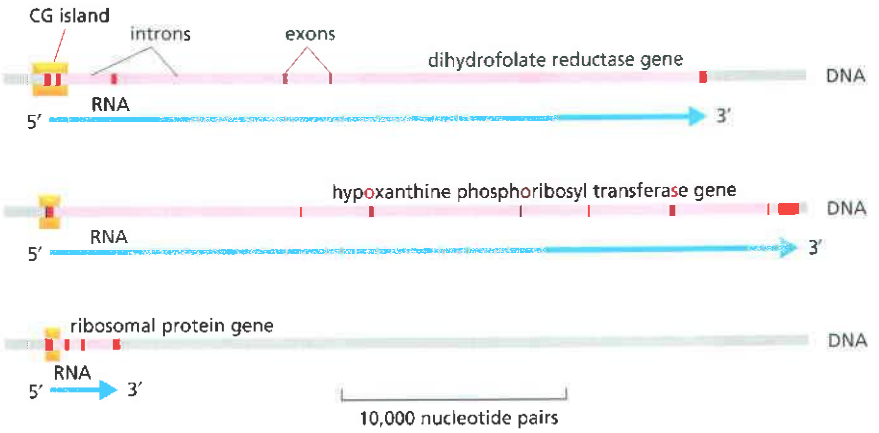


Figure 7–84 The CG islands surrounding the promoter in three mammalian housekeeping genes. The *yellow boxes* show the extent of each island. As for most genes in mammals (see Figure 6–25), the exons *(dark red)* are very short relative to the introns *(light red)*. (Adapted from A.P. Bird, *Trends Genet.* 3:342–347, 1987. With permission from Elsevier.)

periods of evolutionary time, the methylated CG sequences in these inactive regions have presumably been lost through spontaneous deamination events that were not properly repaired. However, promoters of genes that remain active in the germ cell lineages (including most housekeeping genes) are kept unmethylated, and therefore spontaneous deaminations of Cs that occur within them can be accurately repaired. Such regions are preserved in modern-day vertebrate cells as CG islands (**Figure 7–85**). In addition, any mutation of a CG sequence in the genome that destroyed the function or regulation of a gene in the adult would be selected against, and some CG islands are presumably the result of a higher than normal density of critical CG sequences for these genes.

The mammalian genome contains an estimated 20,000 CG islands. Most of the islands mark the 5′ ends of transcription units and thus, presumably, of genes. The presence of CG islands thereby provides a convenient way of identifying genes in the DNA sequences of vertebrate genomes.

Epigenetic Mechanisms Ensure That Stable Patterns of Gene Expression Can Be Transmitted to Daughter Cells

As we have seen, once a cell in an organism differentiates into a particular cell type, it generally remains specialized in that way; if it divides, its daughters inherit the same specialized character. For example, liver cells, pigment cells, and endothelial cells (discussed in Chapter 23) divide many times in the life of an individual, each of them faithfully producing daughter cells of the same type. Such differentiated cells must remember their specific pattern of gene expression and pass it on to their progeny through all subsequent cell divisions.

We have already described several ways of enabling daughter cells to "remember" what kind of cells they are supposed to be. One of the simplest is through a positive feedback loop in which a key gene regulatory protein activates, either directly or indirectly, the transcription of its own gene (see Figures 7–68 and 7–69). Interlocking positive feedback loops provide even more stability by buffering the circuit against fluctuations in the level of any one gene regulatory protein (Figures 7–75B and 7–78). We also saw above that DNA methylation can serve as a means for propagating gene expression patterns to descendants (see Figure 7–80).

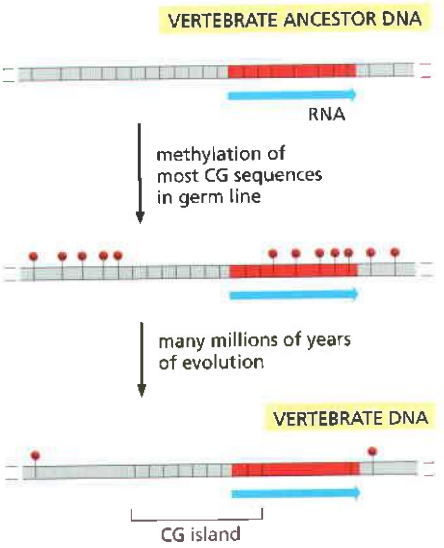


Figure 7–85 A mechanism to explain both the marked overall deficiency of CG sequences and their clustering into CG islands in vertebrate genomes. A *black line* marks the location of a CG dinucleotide in the DNA sequence, while a *red "lollipop"* indicates the presence of a methyl group on the CG dinucleotide. CG sequences that lie in regulatory sequences of genes that are transcribed in germ cells are unmethylated and therefore tend to be retained in evolution. Methylated CG sequences, on the other hand, tend to be lost through deamination of 5-methyl C to T, unless the CG sequence is critical for survival.

Positive feedback loops and DNA methylation are common to both bacteria and eucaryotes; but eucaryotes also have available to them another means of maintaining a differentiated state through many cell generations. As we saw in Chapter 4, chromatin structure itself can be faithfully propagated from parent to daughter cell. There are several mechanisms to bring this about, but the simplest is based on the covalent modifications of histones. As we have seen, these modifications form a "histone code," with different patterns of modification serving as binding sites for different reader proteins. If these proteins, in turn, serve as (or attract) writer enzymes that replicate the very modification patterns that attracted them in the first place, then the distribution of active and silent regions of chromatin can be faithfully propagated (see Figure 5–39). In a sense, self-sustaining modification of histones is a form of positive feedback loop that is tied to the DNA but does not require the participation of the underlying DNA sequences.

The ability of a daughter cell to retain a memory of the gene expression patterns that were present in the parent cell is an example of **epigenetic inheritance**. This term has subtly different meanings in different branches of biology, but we will use it in its broadest sense to cover any heritable difference in the phenotype of a cell or an organism that does not result from changes in the nucleotide sequence of DNA (see Figure 4–35). We have just discussed three of the most important mechanisms underlying epigenetic changes, but others also exist (**Figure 7–86**). Cells often combine these mechanisms to ensure that patterns of gene expression are maintained and inherited accurately and reliably—over a period of up to a hundred years or more, in our own case.

For more than half a century, biologists have been preoccupied with DNA as the carrier of heritable information—and rightly so. However, it has become clear that human chromosomes also carry a great deal of information that is epigenetic, and not contained in the sequence of the DNA itself. Imprinting is one

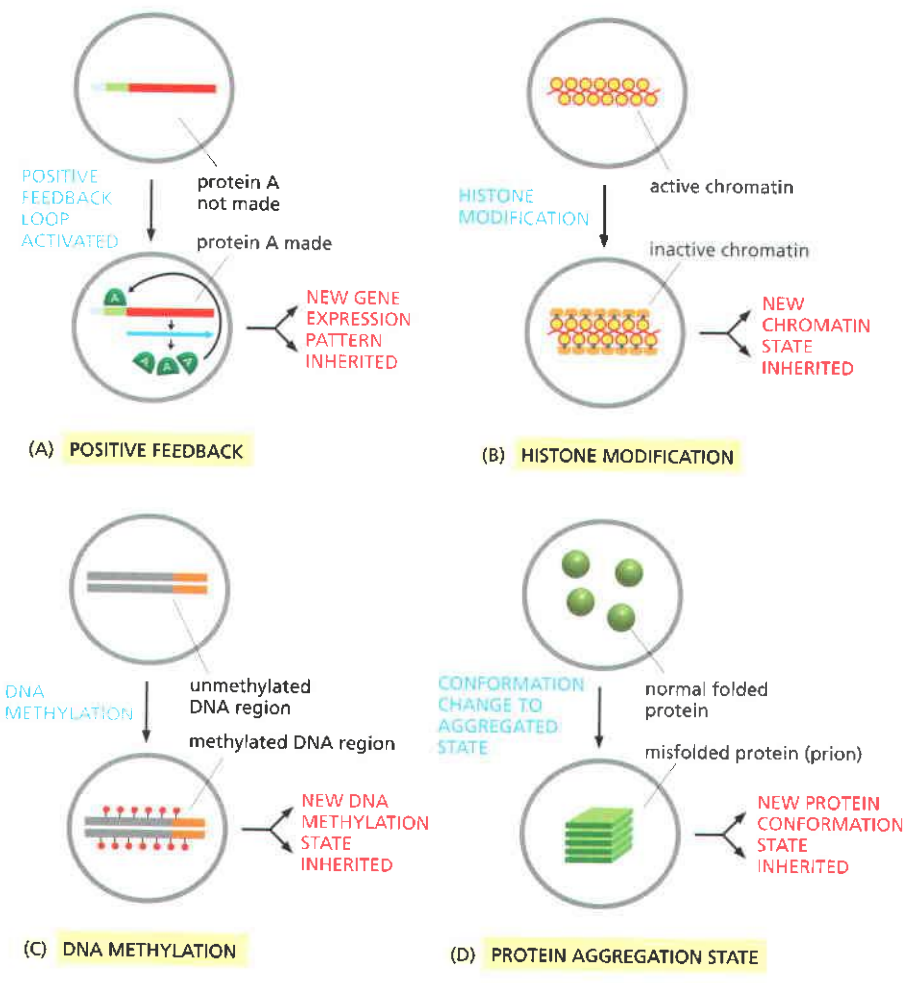


Figure 7–86 Four distinct mechanisms that can produce an epigenetic form of inheritance in an organism. (For the inheritance of histone modifications, see Figure 4–52; for the inheritance of protein conformations, see Figure 6–95.)

example. Another is seen in the phenomenon of *mono-allelic expression*, in which only one of the two copies of certain human genes is expressed. For many such genes, the decision of which allele to express and which to silence is random, but once made, it is passed on to progeny cells. Below, we will see an extreme example of this phenomenon in X-chromosome inactivation.

The net effect of random and environmentally triggered epigenetic changes in humans can be seen by comparing identical twins: their genomes have the same sequence of nucleotides, but when their histone modification and DNA methylation patterns are compared, many differences are observed. Because these differences are roughly correlated not only with age but also with the time that the twins have spent apart from each other, it is believed that some of these changes are the result of environmental factors (**Figure 7–87**). Although studies of the *epigenome* are in early stages, the idea that environmental events can be permanently registered by our cells is a fascinating one that presents an important challenge to the next generation of biological scientists.

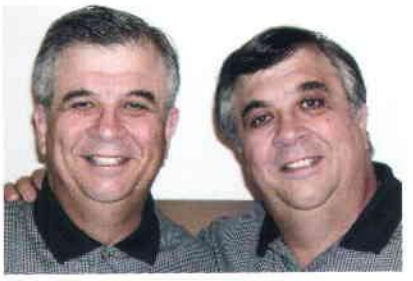


Figure 7–87 Identical twins raised apart from one another. (Courtesy of Nancy L. Segal.)

Chromosome-Wide Alterations in Chromatin Structure Can Be Inherited

We have seen that chromatin states and DNA methylation can be heritable, serving to establish and preserve patterns of gene expression for many cell generations. Perhaps the most striking example of this effect occurs in mammals, in which an alteration in the chromatin structure of an entire chromosome can modulate the levels of expression of all genes on that chromosome.

Males and females differ in their *sex chromosomes*. Females have two X chromosomes, whereas males have one X and one Y chromosome. As a result, female cells contain twice as many copies of X-chromosome genes as do male cells. In mammals, the X and Y sex chromosomes differ radically in gene content: the X chromosome is large and contains more than a thousand genes, whereas the Y chromosome is small and contains less than 100 genes. Mammals have evolved a *dosage compensation* mechanism to equalize the dosage of X-chromosome gene products between males and females. Mutations that interfere with dosage compensation are lethal: the correct ratio of X chromosome to *autosome* (non-sex chromosome) gene products is critical for survival.

Mammals achieve dosage compensation by the transcriptional inactivation of one of the two X chromosomes in female somatic cells, a process known as **X-inactivation**. Early in the development of a female embryo, when it consists of a few thousand cells, one of the two X chromosomes in each cell becomes highly condensed into a type of heterochromatin. The condensed X chromosome can be easily seen under the light microscope in interphase cells; it was originally called a *Barr body* and is located near the nuclear membrane (**Figure 7–88**). As a result of X-inactivation, two X chromosomes can coexist within the same nucleus, exposed to the same diffusible gene regulatory proteins, yet differ entirely in their expression.

The initial choice of which X chromosome to inactivate, the maternally inherited one (X_m) or the paternally inherited one (X_p), is random. Once either X_p or X_m has been inactivated, it remains silent throughout all subsequent cell divisions of that cell and its progeny, indicating that the inactive state is faithfully maintained through many cycles of DNA replication and mitosis. Because X-inactivation is random and takes place after several thousand cells have already formed in the embryo, every female is a mosaic of clonal groups of cells

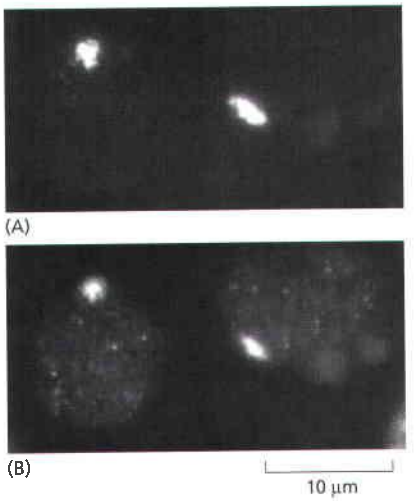


Figure 7–88 X-chromosome inactivation in female cells. (A) Only the inactive X chromosome is coated with XIST RNA, visualized here by *in situ* hybridization to fluorescently labeled RNAs of complementary nucleotide sequence. The panel shows the nuclei of two adjacent cells. (B) The same sample, stained with antibodies against a component of the Polycomb group complex, which coats the X chromosome and helps to silence expression of its genes. (From B. Panning, *Methods Enzymol.* 376:419–428, 2004. With permission from Academic Press.)

in which either X_p or X_m is silenced (**Figure 7–89**). These clonal groups are distributed in small clusters in the adult animal because sister cells tend to remain close together during later stages of development. For example, X-chromosome inactivation causes the red and black "tortoise-shell" coat coloration of some female cats. In these cats, one X chromosome carries a gene that produces red hair color, and the other X chromosome carries an allele of the same gene that results in black hair color; it is the random X-inactivation that produces patches of cells of two distinctive colors. In contrast to the females, male cats of this genetic stock are either solid red or solid black, depending on which X chromosome they inherit from their mothers.

Although X-chromosome inactivation is maintained over thousands of cell divisions, it is not always permanent. In particular, it is reversed during germ-cell formation, so that all haploid oocytes contain an active X chromosome and can express X-linked gene products.

How is an entire chromosome transcriptionally inactivated? X-chromosome inactivation is initiated and spreads from a single site in the middle of the X chromosome, the **X-inactivation center** (**XIC**). Encoded within the XIC is an unusual RNA molecule, *XIST RNA*, which is expressed solely from the inactive X chromosome and whose expression is necessary for X-inactivation. The XIST RNA is not translated into protein and remains in the nucleus, where it eventually coats the entire inactive X chromosome. The spread of XIST RNA from the XIC over the entire chromosome correlates with the spread of gene silencing, indicating that XIST RNA drives the formation and spread of heterochromatin (**Figure 7–90**). Curiously, about 10% of the genes on the X-chromosome escape this silencing and remain active.

In addition to containing XIST RNA, the X-chromosome heterochromatin is characterized by a specific variant of histone 2A, by hypoacetylation of histones

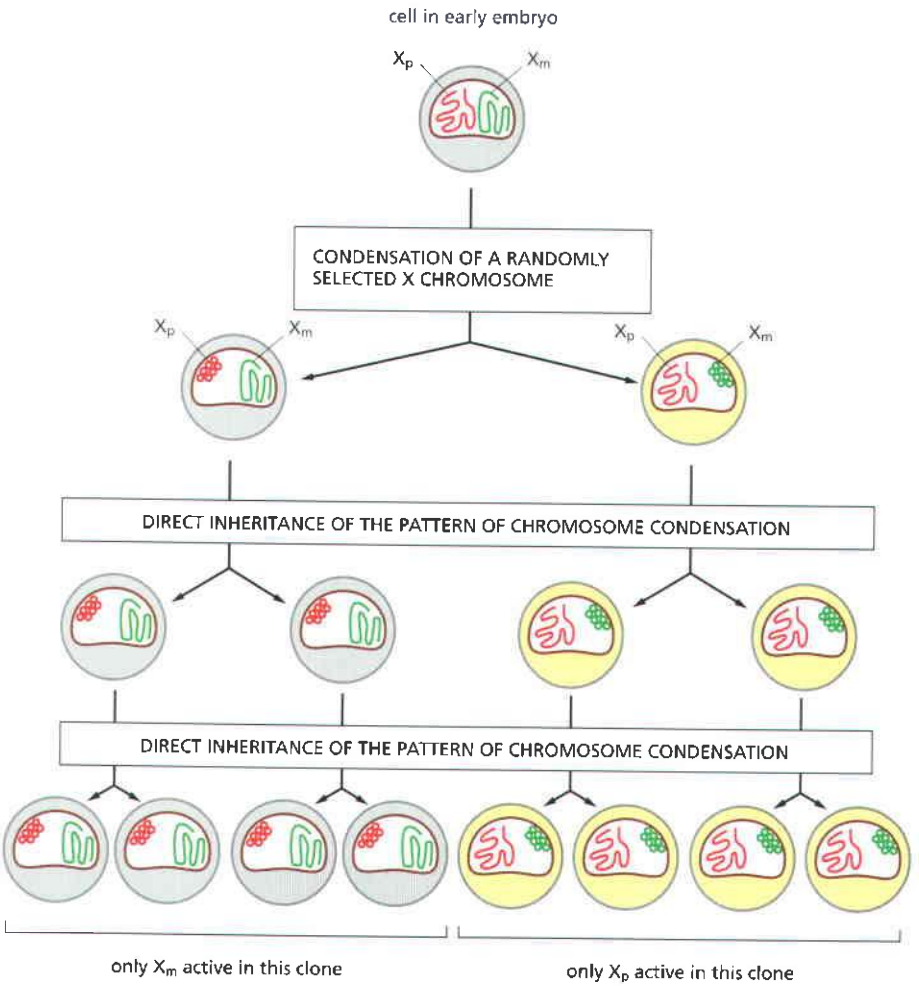


Figure 7–89 X-inactivation. The clonal inheritance of a condensed inactive X chromosome that occurs in female mammals.

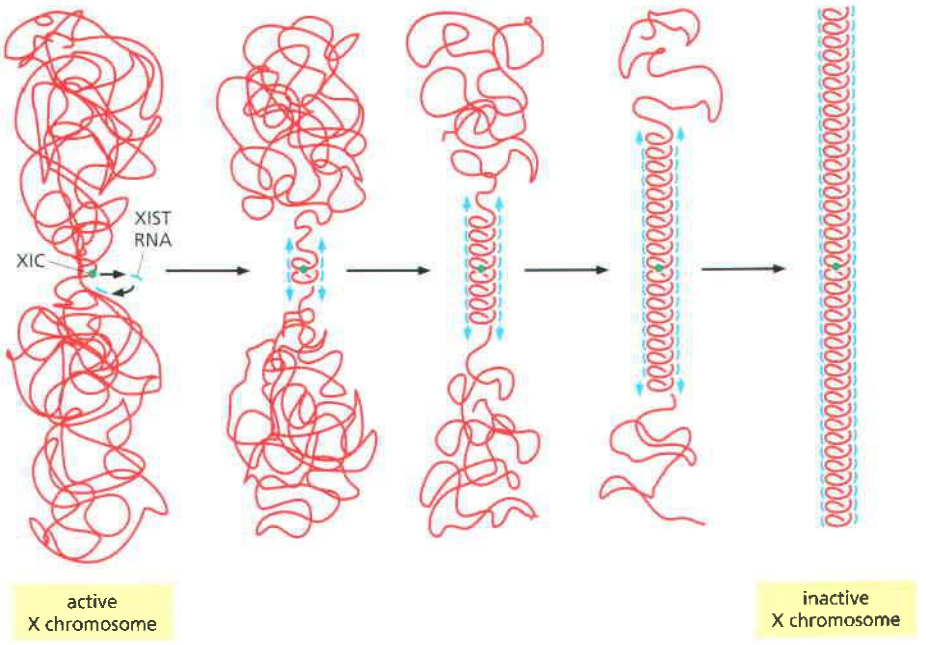


Figure 7–90 Mammalian X-chromosome inactivation. X-chromosome inactivation begins with the synthesis of XIST (X-inactivation specific transcript) RNA from the XIC (X-inactivation center) locus. The association of XIST RNA with one of a female's two X chromosomes is correlated with the condensation of that chromosome. Early in embryogenesis, both XIST association and chromosome condensation gradually move from the XIC locus outward to the chromosome ends. The details of how this occurs remain to be deciphered.

H3 and H4, by ubiquitylation of histone 2A, by methylation of a specific position on histone H3, and by specific patterns of methylation on the underlying DNA (for a suggestion of how these features may be causally linked, see Figure 7–81). The combination of such modifications presumably makes most of the inactive X chromosome unusually resistant to transcription. Because these modifications are, at least in principle, self-propagating, it is easy to see how, once formed, an inactive X chromosome can be stably maintained through many cell divisions.

Many features of mammalian X-chromosome inactivation remain to be discovered. How is the initial decision made as to which X chromosome to inactivate? What mechanism prevents the other X chromosome from also being inactivated? How does XIST RNA coordinate the formation of heterochromatin? How do some genes on the X chromosome escape inactivation? We are just beginning to understand this mechanism of gene regulation that is crucial for the survival of our own species.

X-chromosome inactivation in females is only one of the ways in which sexually reproducing organisms achieve dosage compensation. In *Drosophila*, most of the genes on the single X chromosome present in male cells are transcribed at approximately twofold higher levels than their counterparts in female cells. This male-specific "up-regulation" of transcription results from an alteration in chromatin structure over the entire male X chromosome. A dosage-compensation complex, containing several histone-modifying enzymes as well as two noncoding RNAs transcribed from the X chromosome, assembles at hundreds of positions along the X chromosome and produces patterns of histone modification that are thought to upregulate transcription—through effects on either initiation or elongation—at most genes on the male X chromosome.

The nematode worm uses a third strategy for dosage compensation. Here, the two sexes are male (with one X chromosome) and hermaphrodite (with two X chromosomes), and dosage compensation occurs by an approximately twofold "down-regulation" of transcription from each of the two X chromosomes in cells of the hermaphrodite. This is brought about through chromosome-wide structural changes in the X chromosomes of hermaphrodites (**Figure 7–91**). A dosage-compensation complex, which is completely different from that of *Drosophila* and resembles instead the *condensin* complex that compacts chromosomes during mitosis and meiosis (see Figure 17–27), assembles along each X chromosome of hermaphrodites and, by an unknown mechanism, superimposes a global twofold repression on the normal expression level of each gene.

Although the strategy and components used to cause dosage compensation differ between mammals, flies, and worms, they all involve structural alterations over the entire X chromosome. It seems likely that features of chromosome

structure that are quite general were independently adapted and harnessed during evolution to overcome a highly specific problem in gene regulation encountered by sexually reproducing animals.

The Control of Gene Expression is Intrinsically Noisy

So far in this chapter we have discussed gene expression as though it were a strictly deterministic process, so that, if only one knew the concentrations of all the relevant regulatory proteins and other control molecules, the level of gene expression would be precisely predictable. In reality, there is a large amount of random variation in the behavior of cells. In part, this is because there are random fluctuations in the environment, and these disturb the concentrations of regulatory molecules inside the cell in unpredictable ways. Another possible cause, in some cases, may be chaotic behavior of the intracellular control system: mathematical analysis shows that even quite simple control systems may be acutely sensitive to the control parameters, in such a way that, for example, a tiny difference of initial conditions may lead to a radically different long-term outcome. But in addition to these causes of unpredictability, there is a further, more fundamental reason why all cell behavior is inescapably random to some degree.

Cells are chemical systems consisting of relatively small numbers of molecules, and chemical reactions at the level of individual molecules occur in an essentially random, or *stochastic*, manner. A given molecule has a certain probability per unit time of undergoing a chemical reaction, but whether it will actually do so at any given moment is unpredictable, depending on random thermal collisions and the probabilistic rules of quantum mechanics. The smaller the number of molecules governing a process inside the cell, the more severely it will be affected by the randomness of chemical events at the single-molecule level. Thus there is some degree of randomness in every aspect of cell behavior, but certain processes are liable to be random in the extreme.

The control of transcription, in particular, depends on the precise chemical condition of the gene. Consider a simple idealized case, in which a gene is transcribed so long as it has a transcriptional activator protein bound to its regulatory region, and transcriptionally silent when this protein is not bound. The association/dissociation reaction between the regulatory DNA and the protein is stochastic: if the bound state has a half-life $t_{1/2}$ of an hour, the gene may remain activated sometimes for 30 minutes or less, sometimes for a couple of hours or more at a stretch, before the activator protein dissociates. In this way, transcription will flicker on and off in an essentially random way. The average rate of flickering, and the ratio of the average time spent in the "on" state to the average time spent in the "off" state, will be determined by the k_{off} and k_{on} values for the binding reaction and by the concentration of the activator protein in the cell. The quantity of gene transcripts accumulated in the cell will fluctuate accordingly; if the lifetime of the transcripts is long compared with $t_{1/2}$, the fluctuations will be smoothed out; if it is short, they will be severe.

One way to demonstrate such random fluctuations in the expression of individual gene copies is to genetically engineer cells in which one copy of a gene control region is linked to a sequence coding for a green fluorescent reporter protein, while another copy is linked similarly to a sequence coding for a red fluorescent reporter. Although both these gene constructs are in the same cell and experiencing the same environment, they fluctuate independently in their level of expression. As a result, in a population of cells that all carry the same pair of constructs, some cells appear green, others red, and still others a mixture of the two colors, and thus in varying shades of yellow (see Figure 8–75). More generally, cell fate decisions are often made in a stochastic manner, presumably as a result of such random fluctuations; we shall encounter an example in Chapter 23, where we discuss the genesis of the different types of white blood cells.

In some types of cells, and for some aspects of cell behavior, randomness in the control of gene transcription, such as we have just described, seems to be the major source of random variability; in other cell types, other sources of random variation predominate. Where noise in a control system would be

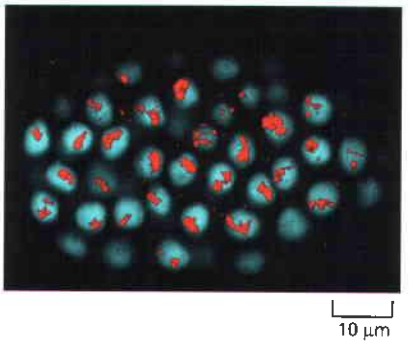


Figure 7–91 Localization of dosage compensation proteins to the X chromosomes of *C. elegans* hermaphrodite (XX) nuclei. This image shows many nuclei from a developing embryo. Total DNA is stained *blue* with the DNA-intercalating dye DAPI, and the Sdc2 protein is stained *red* using anti-Sdc2 antibodies coupled to a fluorescent dye. This experiment shows that the Sdc2 protein associates with only a limited set of chromosomes, identified by other experiments to be the two X chromosomes. Sdc2 is bound along the entire length of the X chromosome and recruits the dosage-compensation complex. (From H.E. Dawes et al., *Science* 284:1800–1804, 1999. With permission from AAAS.)

harmful, special control mechanisms have evolved to minimize its effects; the feed-forward loop discussed earlier is an example of such a device, serving to filter out the effects of rapid fluctuations in a control signal. But in all cells, some degree of randomness is inevitable. It is a fundamental feature of cell behavior.

Summary

The many types of cells in animals and plants are created largely through mechanisms that cause different sets of genes to be transcribed in different cells. Since specialized animal cells can maintain their unique character through many cell division cycles and even when grown in culture, the gene regulatory mechanisms involved in creating them must be stable once established and heritable when the cell divides. These features reflect the cell's memory of its developmental history. Bacteria and yeasts also exhibit cell memory and provide unusually accessible model systems in which to study gene regulatory mechanisms.

Direct or indirect positive feedback loops, which enable gene regulatory proteins to perpetuate their own synthesis, provide the simplest mechanism for cell memory. Transcription circuits also provide the cell with the means to carry out logic operations and measure time. Simple transcription circuits combined into large regulatory networks drive highly sophisticated programs of embryonic development.

In eucaryotes, the transcription of any particular gene is generally controlled by a combination of gene regulatory proteins. It is thought that each type of cell in a higher eucaryotic organism contains a specific set of gene regulatory proteins that ensures the expression of only those genes appropriate to that type of cell. A given gene regulatory protein may be active in a variety of circumstances and is typically involved in the regulation of many different genes.

Unlike bacteria, eucaryotic cells use inherited states of chromatin condensation as an additional mechanism to regulate gene expression and to create cell memory. An especially dramatic case is the inactivation of an entire X chromosome in female mammals. DNA methylation can also silence genes in a heritable manner in eucaryotes. In addition, it underlies the phenomenon of genomic imprinting in mammals, in which the expression of a gene depends on whether it was inherited from the mother or the father.

POST-TRANSCRIPTIONAL CONTROLS

In principle, every step required for the process of gene expression can be controlled. Indeed, one can find examples of each type of regulation, and many genes are regulated by multiple mechanisms. As we have seen, controls on the initiation of gene transcription are a critical form of regulation for all genes. But other controls can act later in the pathway from DNA to protein to modulate the amount of gene product that is made—and in some cases, to determine the exact amino acid sequence of the protein product. These **post-transcriptional controls**, which operate after RNA polymerase has bound to the gene's promoter and begun RNA synthesis, are crucial for the regulation of many genes.

In the following sections, we consider the varieties of post-transcriptional regulation in temporal order, according to the sequence of events that an RNA molecule might experience after its transcription has begun (**Figure 7–92**).

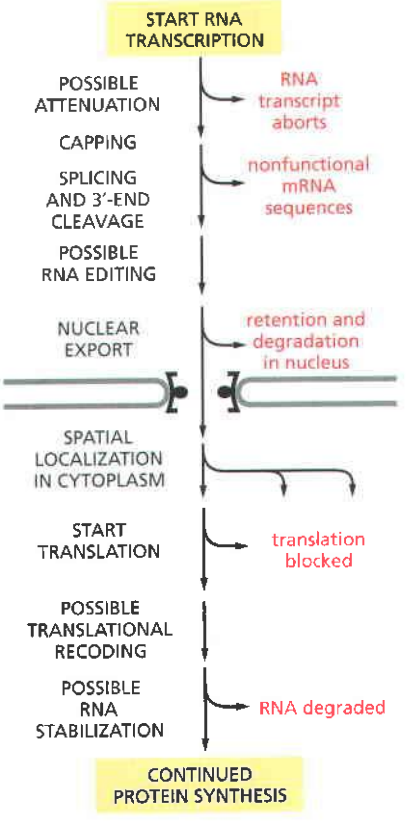


Figure 7–92 Post-transcriptional controls on gene expression. The final synthesis rate of a protein can, in principle, be controlled at any of the steps shown. RNA splicing, RNA editing, and translation recoding (described in Chapter 6) can also alter the sequence of amino acids in a protein, making it possible for the cell to produce more than one protein variant from the same gene. Only a few of the steps depicted here are likely to be critical for the regulation of any one particular protein.

Transcription Attenuation Causes the Premature Termination of Some RNA Molecules

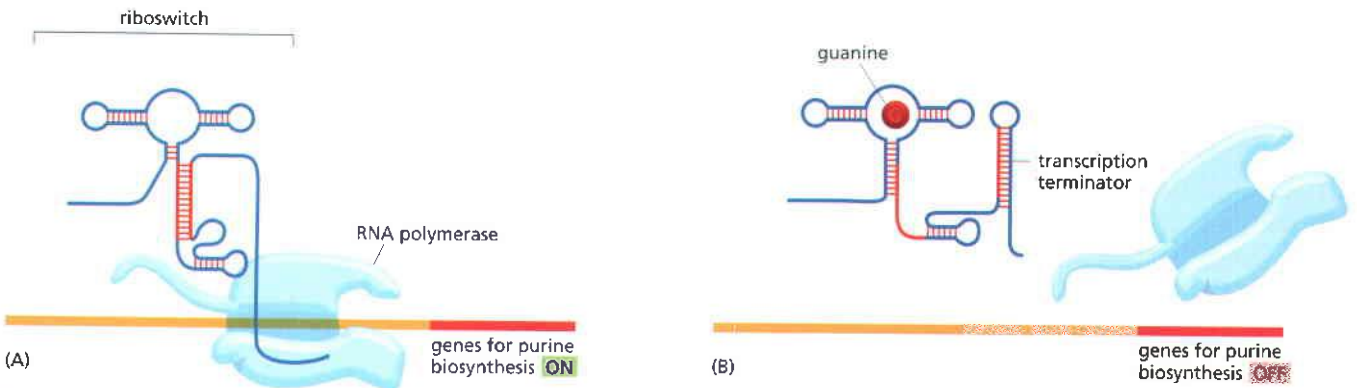
It has long been known that the expression of certain genes in bacteria is inhibited by premature termination of transcription, a phenomenon called **transcription attenuation**. In some of these cases the nascent RNA chain adopts a structure that causes it to interact with the RNA polymerase in such a way as to abort its transcription. When the gene product is required, regulatory proteins bind to the nascent RNA chain and interfere with attenuation, allowing the transcription of a complete RNA molecule.

Transcription attenuation also operates in eucaryotes. A well-studied example occurs during the life cycle of HIV, the human immunodeficiency virus, causative agent of acquired immune deficiency syndrome, or AIDS. Once it has been integrated into the host genome, the viral DNA is transcribed by the cell's RNA polymerase II (see Figure 5–71). However, the host polymerase usually terminates transcription after synthesizing transcripts of several hundred nucleotides and therefore does not efficiently transcribe the entire viral genome. When conditions for viral growth are optimal, a virus-encoded protein called Tat, which binds to a specific stem-loop structure in the nascent RNA that contains a "bulged base," prevents this premature termination. Once bound to this specific RNA structure (called Tar), Tat assembles several cell proteins that allow the RNA polymerase to continue transcribing. The normal role of at least some of these proteins is to prevent pausing and premature termination by RNA polymerase when it transcribes normal cell genes. Eucaryotic genes often contain long introns; to transcribe a gene efficiently, RNA polymerase II cannot afford to linger at nucleotide sequences that happen to promote pausing. Thus, a normal cell mechanism has apparently been adapted by HIV to permit efficient transcription of its genome to be controlled by a single viral protein.

Riboswitches Might Represent Ancient Forms of Gene Control

In Chapter 6, we discussed the idea that, before modern cells arose on Earth, RNA both stored hereditary information and catalyzed chemical reactions. The recent discovery of *riboswitches* shows that RNA can also form control devices that regulate gene expression. Riboswitches are short sequences of RNA that change their conformation on binding small molecules, such as metabolites. Each riboswitch recognizes a specific small molecule and the resulting conformational change is used to regulate gene expression. Riboswitches are often located near the 5′ end of mRNAs, and they fold while the mRNA is being synthesized blocking or permitting progress of the RNA polymerase according to whether the regulatory small molecule is bound (**Figure 7–93**).

Riboswitches are particularly common in bacteria, in which they sense key small metabolites in the cell and adjust gene expression accordingly. Perhaps their most remarkable feature is the high specificity and affinity with which each recognizes only the appropriate small molecule; in many cases, every chemical feature of the small molecule is read by the RNA (Figure 7–93C). Moreover, the



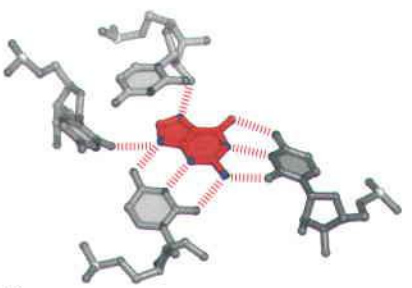

(C)

Figure 7–93 A riboswitch that responds to guanine. (A) In this example from bacteria, the riboswitch controls expression of the purine biosynthetic genes. When guanine levels in cells are low, an elongating RNA polymerase transcribes the purine biosynthetic genes, and the enzymes needed for guanine synthesis are therefore expressed. (B) When guanine is abundant, it binds the riboswitch, causing it to undergo a conformational change that forces the RNA polymerase to terminate transcription (see Figure 6–11). (C) Guanine *(red)* bound to the riboswitch. Only those nucleotides that form the guanine-binding pocket are shown. Many other riboswitches exist, including those that recognize *S*-adenosyl methione, coenzyme B_{12}, flavin mononucleotide, adenine, lysine, and glycine. (Adapted from M. Mandal and R.R. Breaker, *Nat. Rev. Mol. Cell Biol.* 5:451–63, 2004, with permission from Macmillan Publishers Ltd., and C.K. Vanderpool and S. Gottesman, *Mol. Microbiol.* 54:1076–1089, 2004, with permission from Blackwell Publishing.)

binding affinities observed are as tight as those typically observed between small molecules and proteins.

Riboswitches are perhaps the most economical examples of gene control devices, inasmuch as they bypass the need for regulatory proteins altogether. In the example shown in Figure 7–93, the riboswitch controls transcription elongation, but they also regulate other steps in gene expression, as we shall see later in this chapter. Clearly, highly sophisticated gene control devices can be made from short sequences of RNA.

Alternative RNA Splicing Can Produce Different Forms of a Protein from the Same Gene

As discussed in Chapter 6, RNA splicing shortens the transcripts of many eucaryotic genes by removing the intron sequences from the mRNA precursor. We also saw that a cell can splice an RNA transcript differently and thereby make different polypeptide chains from the same gene—a process called **alternative RNA splicing** (see Figure 6–27 and **Figure 7–94**). A substantial proportion of animal genes (estimated at 40% in flies and 75% in humans) produce multiple proteins in this way.

When different splicing possibilities exist at several positions in the transcript, a single gene can produce dozens of different proteins. In one extreme case, a *Drosophila* gene may produce as many as 38,000 different proteins from a single gene through alternative splicing (**Figure 7–95**), although only a fraction of these forms have thus far been experimentally observed. Considering that the *Drosophila* genome has approximately 14,000 identified genes, it is clear that the protein complexity of an organism can greatly exceed the number of its genes. This example also illustrates the perils in equating gene number with an organism's complexity. For example, alternative splicing is relatively rare in single-celled budding yeasts but very common in flies. Budding yeast has ~6200 genes, only about 300 of which are subject to splicing, and nearly all of these have only a single intron. To say that flies have only 2–3 times as many genes as yeasts greatly underestimates the difference in complexity of these two genomes.

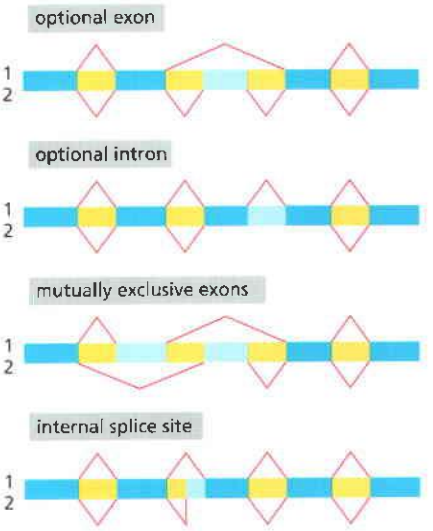


Figure 7–94 Four patterns of alternative RNA splicing. In each case a single type of RNA transcript is spliced in two alternative ways to produce two distinct mRNAs (1 and 2). The *dark blue boxes* mark exon sequences that are retained in both mRNAs. The *light blue boxes* mark possible exon sequences that are included in only one of the mRNAs. The boxes are joined by *red lines* to indicate where intron sequences *(yellow)* are removed. (Adapted with permission from A. Andreadis, M.E. Gallego and B. Nadal-Ginard, *Annu. Rev. Cell Biol.* 3:207–242, 1987. With permission from Annual Reviews.)

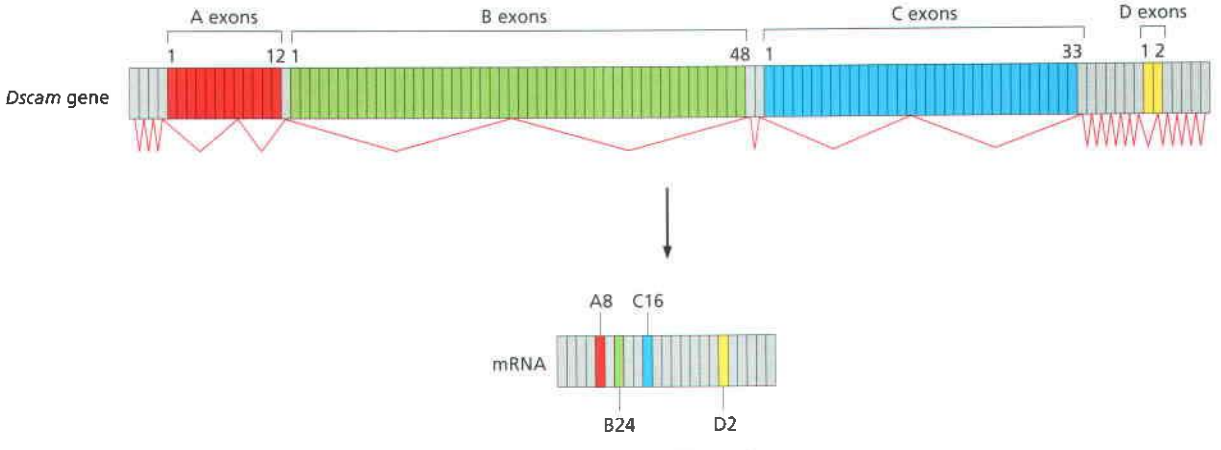


Figure 7–95 Alternative splicing of RNA transcripts of the *Drosophila Dscam* gene. DSCAM proteins are axon guidance receptors that help to direct growth cones to their appropriate targets in the developing nervous system. The final mRNA contains 24 exons, four of which (denoted A, B, C, and D) are present in the *Dscam* gene as arrays of alternative exons. Each RNA contains 1 of 12 alternatives for exon A *(red)*, 1 of 48 alternatives for exon B *(green)*, 1 of 33 alternatives for exon C *(blue)*, and 1 of 2 alternatives for exon D *(yellow)*. If all possible splicing combinations are used, 38,016 different proteins could in principle be produced from the *Dscam* gene. This figure shows only one of the many possible splicing patterns (indicated by the red line and by the mature mRNA below it). Each variant Dscam protein would fold into roughly the same structure [predominantly a series of extracellular immunoglobulin-like domains linked to a membrane-spanning region (see Figure 25–74)], but the amino acid sequence of the domains would vary according to the splicing pattern. It is suspected that this receptor diversity contributes to the formation of complex neural circuits, but the precise properties and functions of the many Dscam variants are not yet understood. (Adapted from D.L. Black, *Cell* 103:367–370, 2000. With permission from Elsevier.)

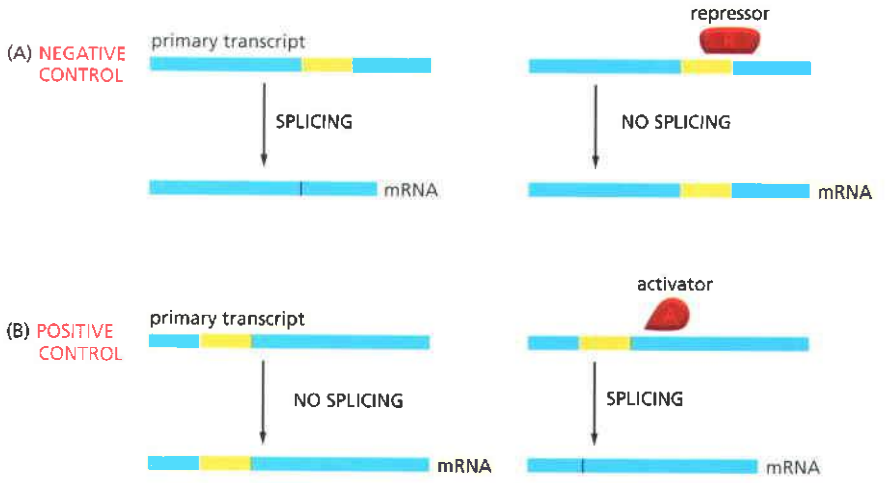


Figure 7–96 Negative and positive control of alternative RNA splicing. (A) In negative control, a repressor protein binds to the pre-mRNA transcript and blocks access of the splicing machinery to a splice junction. This often results in the use of a cryptic splice site, thereby producing an altered pattern of splicing (not shown). (B) In positive control, the splicing machinery is unable to remove a particular intron sequence efficiently without assistance from an activator protein. Because the nucleotide sequences that bind these activators can be located many nucleotide pairs from the splice junctions they control, they are often called *splicing enhancers*.

In some cases alternative RNA splicing occurs because there is an *intron sequence ambiguity*: the standard spliceosome mechanism for removing intron sequences (discussed in Chapter 6) is unable to distinguish cleanly between two or more alternative pairings of 5′ and 3′ splice sites, so that different choices are made by chance on different transcripts. Where such constitutive alternative splicing occurs, several versions of the protein encoded by the gene are made in all cells in which the gene is expressed.

In many cases, however, alternative RNA splicing is regulated rather than constitutive. In the simplest examples, regulated splicing is used to switch from the production of a nonfunctional protein to the production of a functional one. The transposase that catalyzes the transposition of the *Drosophila* P element, for example, is produced in a functional form in germ cells and a nonfunctional form in somatic cells of the fly, allowing the P element to spread throughout the genome of the fly without causing damage in somatic cells (see Figure 5–69). The difference in transposon activity has been traced to the presence of an intron sequence in the transposase RNA that is removed only in germ cells.

In addition to switching from the production of a functional protein to the production of a nonfunctional one, the regulation of RNA splicing can generate different versions of a protein in different cell types, according to the needs of the cell. Tropomyosin, for example, is produced in specialized forms in different types of cells (see Figure 6–27). Cell-type-specific forms of many other proteins are produced in the same way.

RNA splicing can be regulated either negatively, by a regulatory molecule that prevents the splicing machinery from gaining access to a particular splice site on the RNA, or positively, by a regulatory molecule that helps direct the splicing machinery to an otherwise overlooked splice site (**Figure 7–96**).

Because of the plasticity of RNA splicing, the blocking of a "strong" splicing site will often expose a "weak" site and result in a different pattern of splicing. Likewise, activating a suboptimal splice site can result in alternative splicing by suppressing a competing splice site. Thus the splicing of a pre-mRNA molecule can be thought of as a delicate balance between competing splice sites—a balance that can easily be tipped by regulatory proteins.

The Definition of a Gene Has Had to Be Modified Since the Discovery of Alternative RNA Splicing

The discovery that eucaryotic genes usually contain introns and that their coding sequences can be assembled in more than one way raised new questions about the definition of a gene. A gene was first clearly defined in molecular terms in the early 1940s from work on the biochemical genetics of the fungus *Neurospora*. Until then, a gene had been defined operationally as a region of the genome that segregates as a single unit during meiosis and gives rise to a definable phenotypic trait, such as a red or a white eye in *Drosophila* or a round

or wrinkled seed in peas. The work on *Neurospora* showed that most genes correspond to a region of the genome that directs the synthesis of a single enzyme. This led to the hypothesis that one gene encodes one polypeptide chain. The hypothesis proved fruitful for subsequent research; as more was learned about the mechanism of gene expression in the 1960s, a gene became identified as that stretch of DNA that was transcribed into the RNA coding for a single polypeptide chain (or a single structural RNA such as a tRNA or an rRNA molecule). The discovery of split genes and introns in the late 1970s could be readily accommodated by the original definition of a gene, provided that a single polypeptide chain was specified by the RNA transcribed from any one DNA sequence. But it is now clear that many DNA sequences in higher eucaryotic cells can produce a set of distinct (but related) proteins by means of alternative RNA splicing. How, then, is a gene to be defined?

In those relatively rare cases in which a single transcription unit produces two very different eucaryotic proteins, the two proteins are considered to be produced by distinct genes that overlap on the chromosome. It seems unnecessarily complex, however, to consider most of the protein variants produced by alternative RNA splicing as being derived from overlapping genes. A more sensible alternative is to modify the original definition to count as a gene any DNA sequence that is transcribed as a single unit and encodes one set of closely related polypeptide chains (protein isoforms). This definition of a gene also accommodates those DNA sequences that encode protein variants produced by post-transcriptional processes other than RNA splicing, such as translational frameshifting (see Figure 6–78), regulated poly-A addition, and RNA editing (to be discussed below).

Sex Determination in *Drosophila* Depends on a Regulated Series of RNA Splicing Events

We now turn to one of the best-understood examples of regulated RNA splicing. In *Drosophila* the primary signal for determining whether the fly develops as a male or female is the ratio of the number of X chromosomes (X) to the number of autosomal sets (A). Individuals with an X/A ratio of 1 (normally two X chromosomes and two sets of autosomes) develop as females, whereas those with an X/A a ratio of 0.5 (normally one X chromosome and two sets of autosomes) develop as males. This ratio is assessed early in development and is remembered thereafter by each cell. Three crucial gene products transmit information about this ratio to the many other genes that specify male and female characteristics (**Figure 7–97**). As explained in **Figure 7–98**, sex determination in *Drosophila* depends on a cascade of regulated RNA splicing events that involves these three gene products.

Although *Drosophila* sex determination provides one of the best-understood examples of a regulatory cascade based on RNA splicing, it is not clear why the fly should use this strategy. Other organisms (the nematode, for example) use an entirely different scheme for sex determination—one based on transcriptional and translational controls. Moreover, the *Drosophila* male-determination pathway requires that a number of nonfunctional RNA molecules be continually produced, which seems unnecessarily wasteful. One speculation is that this RNA-splicing cascade, like the ribsoswitches discussed above, represents an ancient control strategy, left over from the early stage of evolution in which RNA

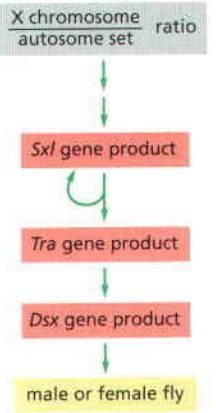


Figure 7–97 Sex determination in *Drosophila*. The gene products shown act in a sequential cascade to determine the sex of the fly according to the X-chromosome/autosome set ratio (X/A). The genes are called *Sex-lethal (Sxl)*, *Transformer (Tra)*, and *Doublesex (Dsx)* because of the phenotypes that result when the gene is inactivated by mutation. The function of these gene products is to transmit the information about the X/A ratio to the many other genes that create the sex-related phenotypes. These other genes function as two alternative sets: those that specify female features and those that specify male features (see Figure 7–98).

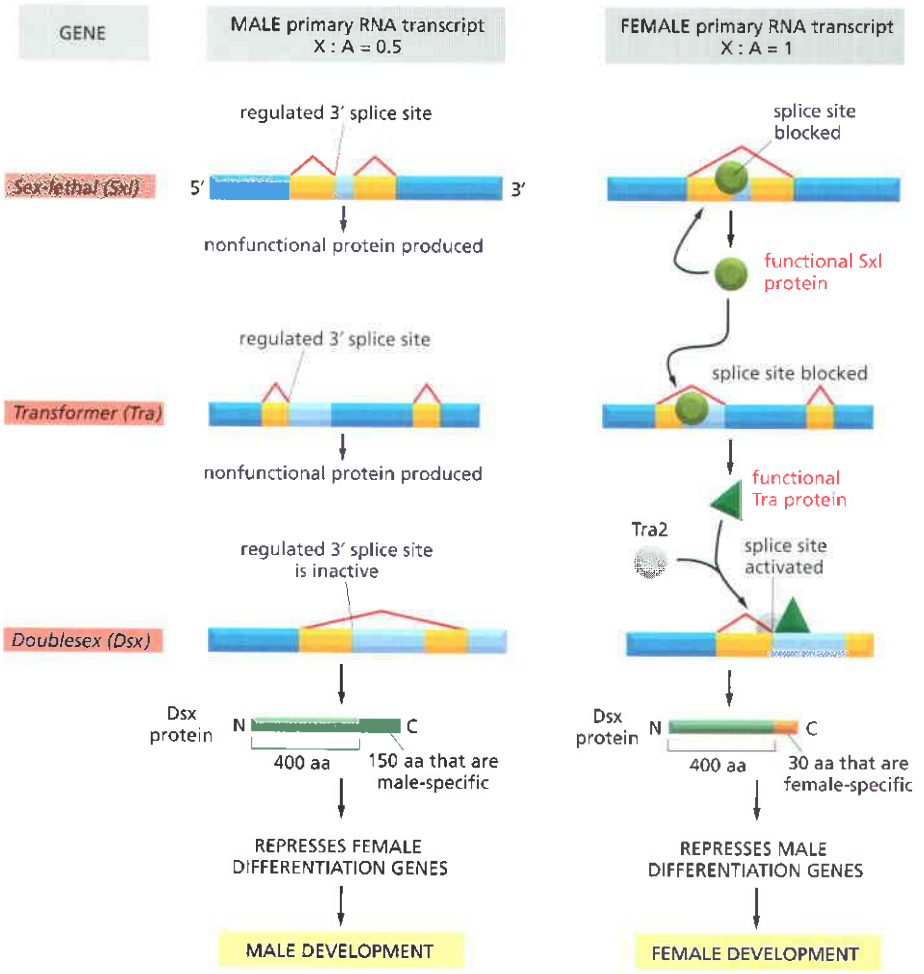


Figure 7–98 The cascade of changes in gene expression that determines the sex of a fly through alternative RNA splicing. An X-chromosome/autosome set ratio of 0.5 results in male development. Male is the "default" pathway in which the *Sxl* and *Tra* genes are both transcribed, but the RNAs are spliced constitutively to produce only nonfunctional RNA molecules, and the *Dsx* transcript is spliced to produce a protein that turns off the genes that specify female characteristics. An X/A ratio of 1 triggers the female differentiation pathway in the embryo by transiently activating a promoter within the *Sxl* gene that causes the synthesis of a special class of *Sxl* transcripts that are constitutively spliced to give functional Sxl protein. Sxl is a splicing regulatory protein with two sites of action: (1) it binds to the constitutively produced *Sxl* RNA transcript, causing a female-specific splice that continues the production of a functional Sxl protein, and (2) it binds to the constitutively produced *Tra* RNA and causes an alternative splice of this transcript, which now produces an active Tra regulatory protein. The Tra protein acts with the constitutively produced Tra2 protein to produce the female-specific spliced form of the *Dsx* transcript; this encodes the female form of the Dsx protein, which turns off the genes that specify male features.

The components in this pathway were all initially identified through the study of *Drosophila* mutants that are altered in their sexual development. The *Dsx* gene, for example, derives its name *(Doublesex)* from the observation that a fly lacking this gene product expresses both male-specific and female-specific features. Note that this pathway includes both negative and positive control of splicing (see Figure 7–96). Sxl binds to the pyrimidine-rich stretch of nucleotides that is part of the standard splicing consensus sequence and blocks access by the normal splicing factor, U2AF (see Figure 6–29). Tra binds to specific RNA sequences in an exon and with Tra2 activates a normally suboptimal splicing signal by the binding of U2AF.

was the predominant biological molecule, and controls of gene expression would have had to be based almost entirely on RNA–RNA interactions.

A Change in the Site of RNA Transcript Cleavage and Poly-A Addition Can Change the C-terminus of a Protein

We saw in Chapter 6 that the 3′ end of a eucaryotic mRNA molecule is not formed by the termination of RNA synthesis by the RNA polymerase. Instead, it results from an RNA cleavage reaction that is catalyzed by additional factors while the transcript is elongating (see Figure 6–37). A cell can control the site of this cleavage so as to change the C-terminus of the resultant protein.

A well-studied example is the switch from the synthesis of membrane-bound to secreted antibody molecules that occurs during the development of B lymphocytes (see Figure 25–17). Early in the life history of a B lymphocyte, the antibody it produces is anchored in the plasma membrane, where it serves as a receptor for antigen. Antigen stimulation causes B lymphocytes to multiply and to begin secreting their antibody. The secreted form of the antibody is identical to the membrane-bound form except at the extreme C-terminus. In this part of the protein, the membrane-bound form has a long string of hydrophobic amino acids that traverses the lipid bilayer of the membrane, whereas the secreted form has a much shorter string of hydrophilic amino acids. The switch from membrane-bound to secreted antibody therefore requires a different nucleotide sequence at the 3′ end of the mRNA; this difference is generated through a change in the length of the primary RNA transcript, caused by a change in the site of RNA cleavage, as shown in **Figure 7–99**. This change is caused by an increase in the concentration of a subunit of CstF, the protein that binds to the G/U-rich sequences of RNA cleavage and poly-A addition sites and promotes

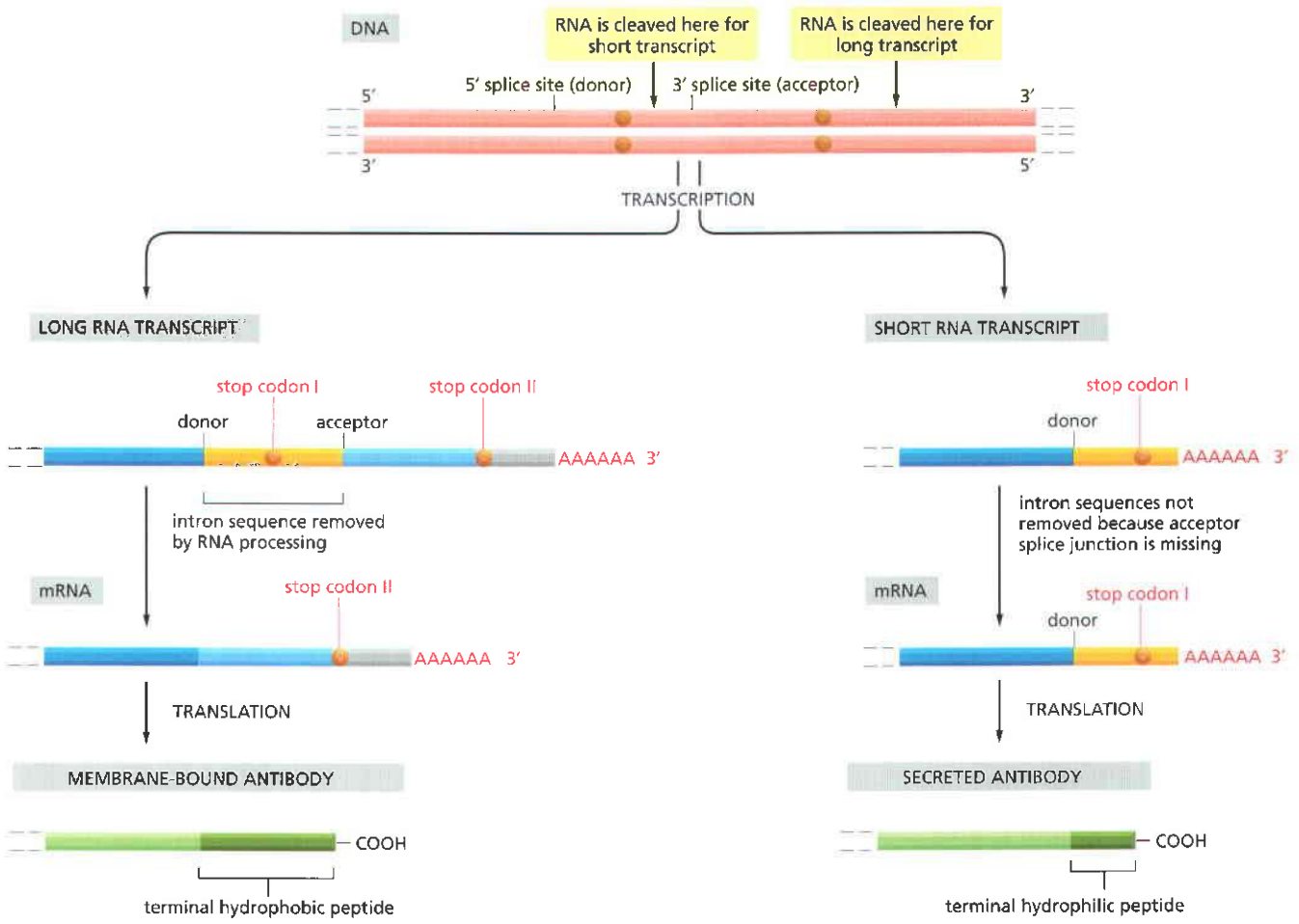


Figure 7–99 Regulation of the site of RNA cleavage and poly-A addition determines whether an antibody molecule is secreted or remains membrane-bound. In unstimulated B lymphocytes *(left)*, a long RNA transcript is produced, and the intron sequence near its 3′ end is removed by RNA splicing to give rise to an mRNA molecule that codes for a membrane-bound antibody molecule. In contrast, after antigen stimulation *(right)*, the primary RNA transcript is cleaved upstream from the splice site in front of the last exon sequence. As a result, some of the intron sequence that is removed from the long transcript remains as a coding sequence in the short transcript. These are the nucleotide sequences that encode the hydrophilic C-terminal portion of the secreted antibody molecule.

RNA cleavage (see Figures 6–37 and 6–38). The first cleavage-poly-A addition site that an RNA polymerase transcribing the antibody gene encounters is suboptimal and is usually skipped in unstimulated B lymphocytes, leading to production of the longer RNA transcript. When activated to produce antibodies, the B lymphocyte increases its CstF concentration; as a result, cleavage now occurs at the suboptimal site, and the shorter transcript is produced. In this way, a change in concentration of a general RNA processing factor can have a dramatic effect on the expression of a particular gene.

RNA Editing Can Change the Meaning of the RNA Message

The molecular mechanisms used by cells are a continual source of surprises. An example is the process of **RNA editing**, which alters the nucleotide sequences of RNA transcripts once they are synthesized and thereby changes the coded message they carry. The most dramatic form of RNA editing was discovered in RNA transcripts that code for proteins in the mitochondria of trypanosomes. Here, one or more U nucleotides are inserted (or, less frequently, removed) from selected regions of a transcript, altering both the original reading frame and the sequence and thereby changing the meaning of the message. For some genes the editing is so extensive that over half of the nucleotides in the mature mRNA are U nucleotides that were inserted during the editing process. A set of 40- to 80-nucleotide-long RNA molecules that are transcribed separately contains the information that specifies exactly how the initial RNA transcript is to be altered. These so-called *guide RNAs* have a 5′ end that is complementary in sequence to one end of the region of the transcript to be edited, followed by a sequence that specifies the set of nucleotides to be inserted into the transcript (**Figure 7–100**). The editing mechanism is remarkably complex: at each edited position, the RNA is broken, U nucleotides are added to the broken 3′ end, and the RNA is ligated.

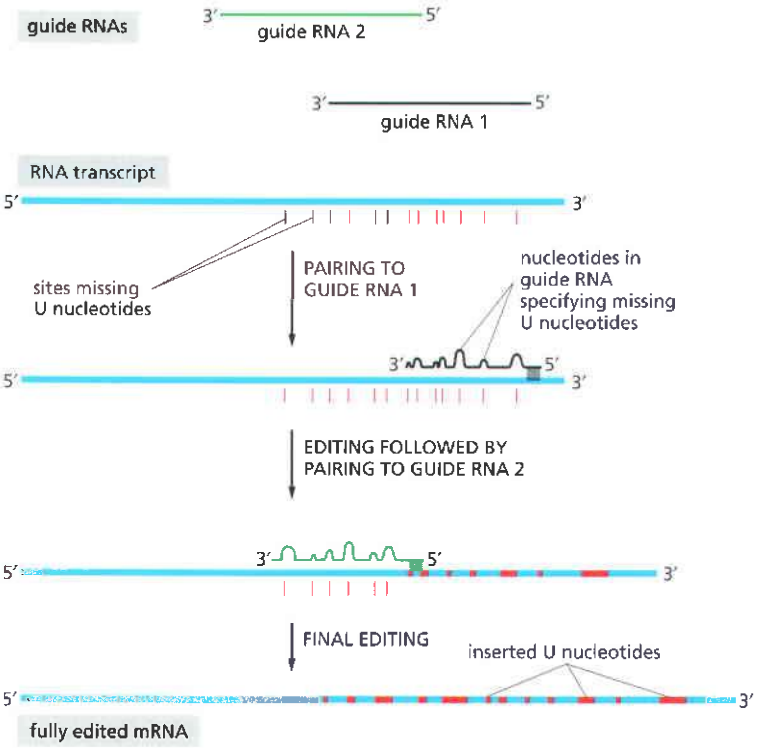


Figure 7–100 RNA editing in the mitochondria of trypanosomes. Editing generally starts near the 3′ end and progresses toward the 5′ end of the RNA transcript, as shown, because the "anchor sequence" at the 5′ end of most guide RNAs can pair only with edited sequences. The U nucleotides are added by a specialized enzyme called a uridylyl transferase.

RNA editing of a more refined type occurs in mammals. Here, two principal types of editing occur, the deamination of adenine to produce inosine (A-to-I editing) and the deamination of cytosine to produce uracil (C-to-U editing; see Figure 5–50). Because these chemical modifications alter the pairing properties of the bases (I pairs with C, and U pairs with A), they can have profound effects on the meaning of the RNA. If the edit occurs in a coding region, it can change the amino acid sequence of the protein or produce a truncated protein. Edits that occur outside coding sequences can affect the pattern of pre-mRNA splicing, the transport of mRNA from the nucleus to the cytosol, or the efficiency with which the RNA is translated.

The process of A-to-I editing is particularly prevalent in humans, where it is estimated to affect over 1000 genes. Enzymes called *ADARs* (adenosine deaminases acting on RNA) perform this type of editing; these enzymes recognize a double-stranded RNA structure that is formed through base pairing between the site to be edited and a complementary sequence located elsewhere on the same RNA molecule, typically in a 3′ intron (**Figure 7–101**). These complementary sequences specify whether the mRNA is to be edited, and if so, exactly where the edit should be made. An especially important example of A-to-I editing takes place in the mRNA that codes for a transmitter-gated ion channel in the brain. A single edit changes a glutamine to an arginine; the affected amino acid lies on the inner wall of the channel, and the editing change alters the Ca^{2+} permeability of the channel. The importance of this edit in mice has been demonstrated by deleting the relevant ADAR gene. The mutant mice are prone to epileptic seizures and die during or shortly after weaning. If the gene for the gated ion channel is mutated to produce the edited form of the protein directly, mice lacking the ADAR develop normally, showing that editing of the ion channel RNA is normally crucial for proper brain development.

C-to-U editing, which is carried out by a different set of enzymes, is also crucial in mammals. For example, in certain cells of the gut, the mRNA for apolipoprotein B undergoes a C-to-U edit that creates a premature stop codon and therefore produces a shorter form of the protein. In cells of the liver, the editing enzyme is not expressed, and the full-length apolipoprotein B is produced. The two protein isoforms have different properties, and each plays a specialized role in lipid metabolism that is specific to the organ that produces it.

Why RNA editing exists at all is a mystery. One idea is that it arose in evolution to correct "mistakes" in the genome. Another is that it arose as a somewhat

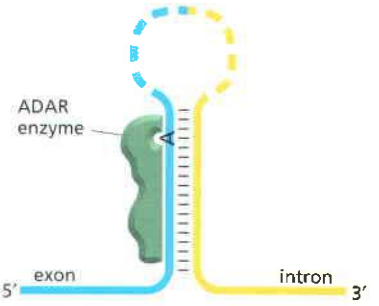


Figure 7–101 Mechanism of A-to-I RNA editing in mammals. RNA sequences carried on the same RNA molecule signal the position of an edit. Typically, a sequence complementary to the position of the edit is present in an intron, and the resulting double-stranded RNA structure attracts the A-to-I editing enzyme ADAR. This type of editing takes place in the nucleus, before the pre-mRNA has been fully processed. Mice and humans have three ADAR enzymes: ADR1 is required in the liver for proper red blood cell development, ADR2 is required for proper brain development (as described in the text), and the role of ADR3 is uncertain.

slapdash way for the cell to produce subtly different proteins from the same gene. A third possibility is that RNA editing originally evolved as a defense mechanism against retroviruses and retrotransposons and was later adapted by the cell to change the meanings of certain mRNAs. Indeed, RNA editing still plays important roles in cell defense. Some retroviruses, including HIV (see Figure 5–71), are extensively edited after they infect cells. This hyperediting creates many harmful mutations in the viral RNA genome and also causes viral mRNAs to be retained in the nucleus, where they are eventually degraded. Although some modern retroviruses protect themselves against this defense mechanism, it presumably helps to hold many viruses in check.

Primates have much higher levels of A-to-I editing than do other mammals, and most of this takes place on RNAs that are transcribed from the highly abundant Alu elements. It has been proposed that A-to-I editing has stopped these mobile elements from completely overtaking our genomes by inactivating the RNA transcripts they require to proliferate (see Figure 5–74). If this idea is correct, RNA editing may have had a profound impact on the shaping of the modern human genome.

RNA Transport from the Nucleus Can Be Regulated

It has been estimated that in mammals only about one-twentieth of the total mass of RNA synthesized ever leaves the nucleus. We saw in Chapter 6 that most mammalian RNA molecules undergo extensive processing and that the "leftover" RNA fragments (excised introns and RNA sequences 3′ to the cleavage/poly-A site) are degraded in the nucleus. Incompletely processed and otherwise damaged RNAs are also eventually degraded as part of the quality control system of RNA production.

As described in Chapter 6, the export of RNA molecules from the nucleus is delayed until processing has been completed. However, mechanisms that deliberately override this control point can be used to regulate gene expression. This strategy forms the basis for one of the best-understood examples of **regulated nuclear transport** of mRNA, which occurs in the human AIDS virus, HIV.

As we saw in Chapter 5, HIV, once inside the cell, directs the formation of a double-stranded DNA copy of its genome, which is then inserted into the genome of the host (see Figure 5–71). Once inserted, the viral DNA is transcribed as one long RNA molecule by the host cell's RNA polymerase II. This transcript is then spliced in many different ways to produce over 30 different species of mRNA, which in turn are translated into a variety of different proteins (**Figure 7–102**). In order to make progeny virus, entire, unspliced viral transcripts must be exported from the nucleus to the cytosol, where they are packaged into viral capsids and serve as the viral genome (see Figure 5–71). This large transcript, as well as alternatively spliced HIV mRNAs that the virus needs to move to the cytoplasm for protein synthesis, still carries complete introns. The host cell's normal block to the nuclear export of unspliced RNAs therefore presents a special problem for HIV.

The block is overcome in an ingenious way. The virus encodes a protein (called Rev) that binds to a specific RNA sequence (called the Rev responsive element, RRE) located within a viral intron. The Rev protein interacts with a nuclear export receptor (exportin 1), which directs the movement of viral RNAs through nuclear pores into the cytosol despite the presence of intron sequences. We discuss in detail the way in which export receptors function in Chapter 12.

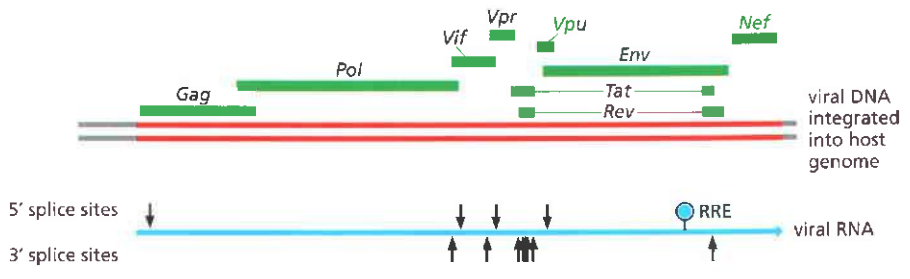


Figure 7–102 The compact genome of HIV, the human AIDS virus. The positions of the nine HIV genes are shown in *green*. The *red double line* indicates a DNA copy of the viral genome that has become integrated into the host DNA *(gray)*. Note that the coding regions of many genes overlap, and that those of *Tat* and *Rev* are split by introns. The *blue line* at the bottom of the figure represents the pre-mRNA transcript of the viral DNA and shows the locations of all the possible splice sites *(arrows)*. There are many alternative ways of splicing the viral transcript; for example the *Env* mRNAs retain the intron that has been spliced out of the *Tat* and *Rev* mRNAs. The Rev response element (RRE) is indicated by a blue ball and stick. It is a 234-nucleotide long stretch of RNA that folds into a defined structure; Rev recognizes a particular hairpin within this larger structure.

The *Gag* gene codes for a protein that is cleaved into several smaller proteins that form the viral capsid. The *Pol* gene codes for a protein that is cleaved to produce reverse transcriptase (which transcribes RNA into DNA), as well as the integrase involved in integrating the viral genome (as double-stranded DNA) into the host genome. Pol is produced by ribosomal frameshifting of translation that begins at *Gag* (see Figure 6–78). The *Env* gene codes for the envelope proteins (see Figure 5–71). Tat, Rev, Vif, Vpr, Vpu, and Nef are small proteins with a variety of functions. For example, Rev regulates nuclear export (see Figure 7–103) and Tat regulates the elongation of transcription across the integrated viral genome (see p. 478).

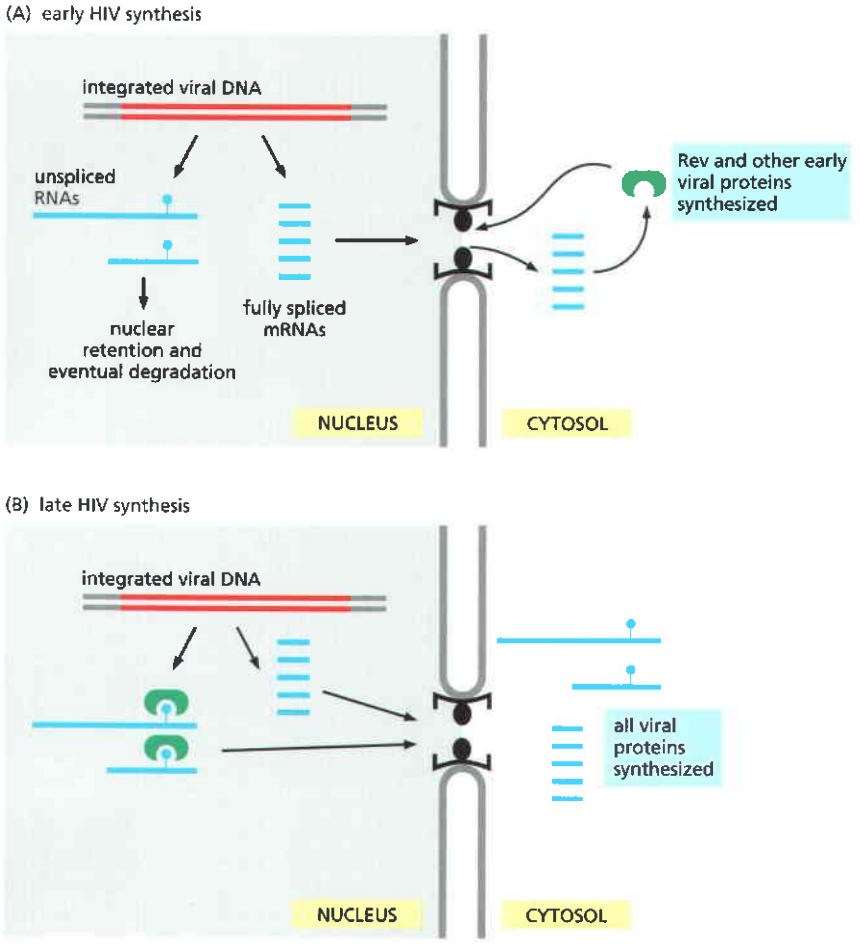


Figure 7–103 Regulation of nuclear export by the HIV Rev protein. Early in HIV infection (A), only the fully spliced RNAs (which contain the coding sequences for Rev, Tat, and Nef) are exported from the nucleus and translated. Once sufficient Rev protein has accumulated and been transported into the nucleus (B), unspliced viral RNAs can be exported from the nucleus. Many of these RNAs are translated into protein, and the full-length transcripts are packaged into new viral particles.

The regulation of nuclear export by Rev has several important consequences for HIV growth and pathogenesis. In addition to ensuring the nuclear export of specific unspliced RNAs, it divides the viral infection into an early phase (in which Rev is translated from a fully spliced RNA and all of intron-containing viral RNAs are retained in the nucleus and degraded) and a late phase (in which unspliced RNAs are exported due to Rev function). This timing helps the virus replicate by providing the gene products in roughly the order in which they are needed (**Figure 7–103**). Regulation by Rev may also help the HIV virus to achieve latency, a condition in which the HIV genome has become integrated into the host cell genome but the production of viral proteins has temporarily ceased. If, after its initial entry into a host cell, conditions became unfavorable for viral transcription and replication, Rev is made at levels too low to promote export of unspliced RNA. This situation stalls the viral growth cycle until conditions improve, Rev levels increase, and the virus enters the replication cycle.

Some mRNAs Are Localized to Specific Regions of the Cytoplasm

Once a newly made eucaryotic mRNA molecule has passed through a nuclear pore and entered the cytosol, it is typically met by ribosomes, which translate it into a polypeptide chain (see Figure 6–40). Once the first round of translation "passes" the nonsense-mediated decay test (see Figure 6–80), the mRNA is usually translated in earnest. If the mRNA encodes a protein that is destined to be secreted or expressed on the cell surface, a signal sequence at the protein's amino terminus will direct it to the endoplasmic reticulum (ER); components of the cell's protein-sorting apparatus recognize the signal sequence as soon as it emerges from the ribosome and direct the entire complex of ribosome, mRNA, and nascent protein to the membrane of the ER, where the remainder of the polypeptide chain is synthesized, as discussed in Chapter 12. In other cases free

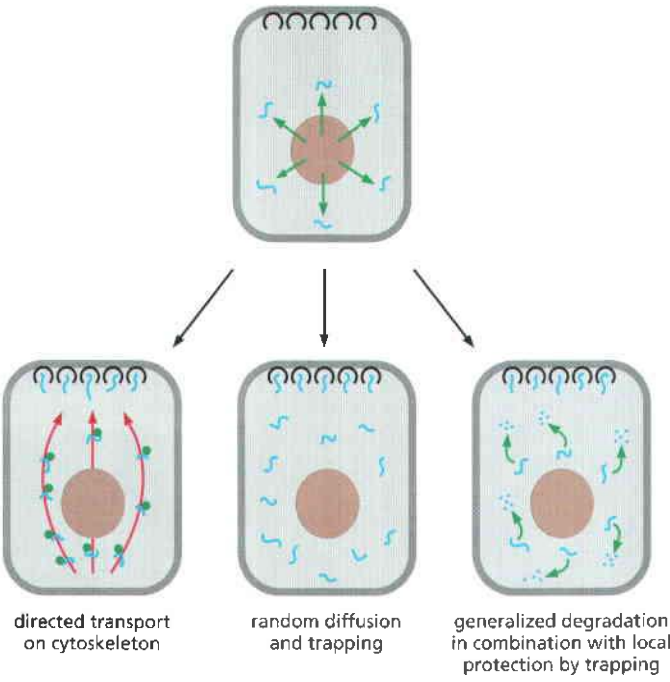


Figure 7–104 Three mechanisms for the localization of mRNAs. The mRNA to be localized leaves the nucleus through nuclear pores *(top)*. Some localized mRNAs *(left diagram)* travel to their destination by associating with cytoskeletal motors. As described in Chapter 16, these motors use the energy of ATP hydrolysis to move unidirectionally along filaments in the cytoskeleton *(red)*. At their destination, anchor proteins *(black)* hold the mRNAs in place. Other mRNAs randomly diffuse through the cytosol and are simply trapped and therefore concentrated at their sites of localization *(center diagram)*. Some of these mRNAs *(right diagram)* are degraded in the cytosol unless they have bound, through random diffusion, a localized protein complex that anchors and protects the mRNA from degradation *(black)*. Each of these mechanisms requires specific signals on the mRNA, which are typically located in the 3′ UTR (see Figure 7–105). In many cases of mRNA localization, additional mechanisms block the translation of the mRNA until it is properly localized. (Adapted from H.D. Lipshitz and C.A. Smibert, *Curr. Opin. Genet. Dev.* 10:476–488, 2000. With permission from Elsevier.)

ribosomes in the cytosol synthesize the entire protein, and signals in the completed polypeptide chain may then direct the protein to other sites in the cell.

Some mRNAs are themselves directed to specific intracellular locations before their efficient translation begins, allowing the cell to position its mRNAs close to the sites where the encoded protein is needed. This strategy provides the cell with many advantages. For example, it allows the establishment of asymmetries in the cytosol of the cell, a key step in many stages of development. Localized mRNA, coupled with translation control, also allows the cell to regulate gene expression independently in its different parts. This feature is particularly important in large, highly polarized cells such as neurons, where growth cones must respond to signals without waiting to involve the distant nucleus. RNA localization has been observed in many organisms, including unicellular fungi, plants, and animals, and it is likely to be a common mechanism that cells use to concentrate high-level production of proteins at specific sites.

Several distinct mechanisms for mRNA localization have been discovered (**Figure 7–104**), but all of them require specific signals in the mRNA itself. These signals are usually concentrated in the 3′ *untranslated region (UTR)*, the region of RNA that extends from the stop codon that terminates protein synthesis to the start of the poly-A tail (**Figure 7–105**). This mRNA localization is usually coupled with translational controls to ensure that the mRNA remains quiescent until it has been moved into place.

The *Drosophila* egg exhibits an especially striking example of mRNA localization. The mRNA encoding the bicoid gene regulatory protein is localized by attachment to the cytoskeleton at the anterior tip of the developing egg. When fertilization triggers the translation of this mRNA, it generates a gradient of the bicoid protein that plays a crucial part in directing the development of the anterior part of the embryo (shown in Figure 7–53 and discussed in more detail in Chapter 22). Many mRNAs in somatic cells are also localized in a similar way. The mRNA that encodes actin, for example, is localized to the actin-filament-rich cell cortex in mammalian fibroblasts by means of a 3′ UTR signal.

We saw in Chapter 6 that mRNA molecules exit from the nucleus bearing numerous markings in the form of RNA modifications (the 5′ cap and the 3′ poly-A tail) and bound proteins (exon-junction complexes, for example) that signify the successful completion of the different pre-mRNA processing steps. As just described, the 3′ UTR of an mRNA can be thought of as a "zip code" that directs mRNAs to different places in the cell. Below we will also see that mRNAs carry information specifying their average lifetime in the cytosol and the efficiency with which they are translated into protein. In a broad sense, the

untranslated regions of eucaryotic mRNAs resemble the transcriptional control regions of genes: their nucleotide sequences contain information specifying the way the RNA is to be used, and proteins interpret this information by binding specifically to these sequences. Thus, over and above the specification of the amino acid sequences of proteins, mRNA molecules are rich with many additional types of information.

The 5′ and 3′ Untranslated Regions of mRNAs Control Their Translation

Once an mRNA has been synthesized, one of the most common ways of regulating the levels of its protein product is to control the step that initiates translation. Even though the details of translation initiation differ between eucaryotes and bacteria (as we saw in Chapter 6), they each use some of the same basic regulatory strategies.

In bacterial mRNAs, a conserved stretch of six nucleotides, the *Shine-Dalgarno sequence*, is always found a few nucleotides upstream of the initiating AUG codon. This sequence forms base pairs with the 16S RNA in the small ribosomal subunit, correctly positioning the initiating AUG codon in the ribosome. Because this interaction strongly contributes to the efficiency of initiation, it provides the bacterial cell with a simple way to regulate protein synthesis through **translational control** mechanisms. These mechanisms, carried out by proteins or by RNA molecules, generally involve either exposing or blocking the Shine–Dalgarno sequence (**Figure 7–106**).

Eucaryotic mRNAs do not contain a Shine–Dalgarno sequence. Instead, as discussed in Chapter 6, the selection of an AUG codon as a translation start site is largely determined by its proximity to the cap at the 5′ end of the mRNA molecule, which is the site at which the small ribosomal subunit binds to the mRNA and begins scanning for an initiating AUG codon. Despite the differences in translation initiation, eucaryotes use similar strategies to regulate translation. For example, translational repressors bind to the 5′ end of the mRNA and thereby inhibit translation initiation. Other repressors recognize nucleotide sequences in the 3′ UTR of specific mRNAs and decrease translation initiation by interfering with the communication between the 5′ cap and 3′ poly-A tail, a step required for efficient translation (see Figure 6–72). A particularly important type of translational control in eucaryotes relies on small RNAs (termed *microRNAs* or *miRNAs*) that bind to mRNAs and reduce protein output. The miRNAs are synthesized and processed in a specialized way, and we shall return to them later in the chapter.

The Phosphorylation of an Initiation Factor Regulates Protein Synthesis Globally

Eucaryotic cells decrease their overall rate of protein synthesis in response to a variety of stressful situations, including deprivation of growth factors or nutrients, infection by viruses, and sudden increases in temperature. Much of this decrease is caused by the phosphorylation of the translation initiation factor eIF2 by specific protein kinases that respond to the changes in conditions.

The normal function of eIF2 was outlined in Chapter 6. It forms a complex with GTP and mediates the binding of the methionyl initiator tRNA to the small ribosomal subunit, which then binds to the 5′ end of the mRNA and begins scanning along the mRNA. When an AUG codon is recognized, the eIF2 protein hydrolyzes the bound GTP to GDP, causing a conformational change in the protein and releasing it from the small ribosomal subunit. The large ribosomal subunit then joins the small one to form a complete ribosome that begins protein synthesis (see Figure 6–71).

Because eIF2 binds very tightly to GDP, a guanine nucleotide exchange factor (see Figure 3–73), designated eIF2B, is required to cause GDP release so that a new GTP molecule can bind and eIF2 can be reused (**Figure 7–107**A). The

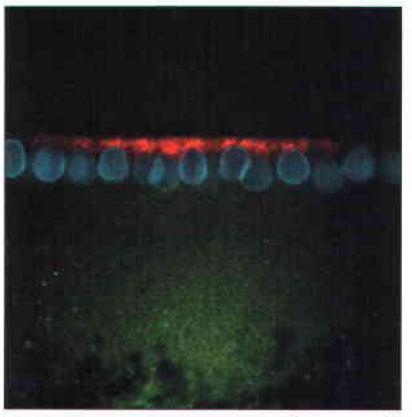


Figure 7–105 An experiment demonstrating the importance of the 3′ UTR in localizing mRNAs to specific regions of the cytoplasm. For this experiment, two different fluorescently labeled RNAs were prepared by transcribing DNA *in vitro* in the presence of fluorescently labeled derivatives of UTP. One RNA (labeled with a *red* fluorochrome) contains the coding region for the *Drosophila* hairy protein and includes the adjacent 3′ UTR (see Figure 6–22). The other RNA (labeled *green*) contains the hairy coding region with the 3′ UTR deleted. The two RNAs were mixed and injected into a *Drosophila* embryo at a stage of development when multiple nuclei reside in a common cytoplasm (see Figure 7–53). When the fluorescent RNAs were visualized 10 minutes later, the full-length hairy RNA *(red)* was localized to the apical side of nuclei *(blue)* but the transcript missing the 3′ UTR *(green)* failed to localize. Hairy is one of many gene regulatory proteins that specifies positional information in the developing *Drosophila* embryo (discussed in Chapter 22). The localization of its mRNA (shown in this experiment to depend on its 3′ UTR) is thought to be critical for proper fly development. (Courtesy of Simon Bullock and David Ish-Horowicz.)

reuse of eIF2 is inhibited when it is phosphorylated—the phosphorylated eIF2 binds to eIF2B unusually tightly, inactivating eIF2B. There is more eIF2 than eIF2B in cells, and even a fraction of phosphorylated eIF2 can trap nearly all of the eIF2B. This prevents the reuse of the nonphosphorylated eIF2 and greatly slows protein synthesis (Figure 7–107B).

Regulation of the level of active eIF2 is especially important in mammalian cells; eIF2 is part of the mechanism that allows cells to enter a nonproliferating, resting state (called G_0)—in which the rate of total protein synthesis is reduced to about one-fifth the rate in proliferating cells (discussed in Chapter 17).

Initiation at AUG Codons Upstream of the Translation Start Can Regulate Eucaryotic Translation Initiation

We saw in Chapter 6 that eucaryotic translation typically begins at the first AUG downstream of the 5′ end of the mRNA, which is the first AUG encountered by a scanning small ribosomal subunit. But the nucleotides immediately surrounding the AUG also influence the efficiency of translation initiation. If the recognition site is poor enough, scanning ribosomal subunits will sometimes ignore the first

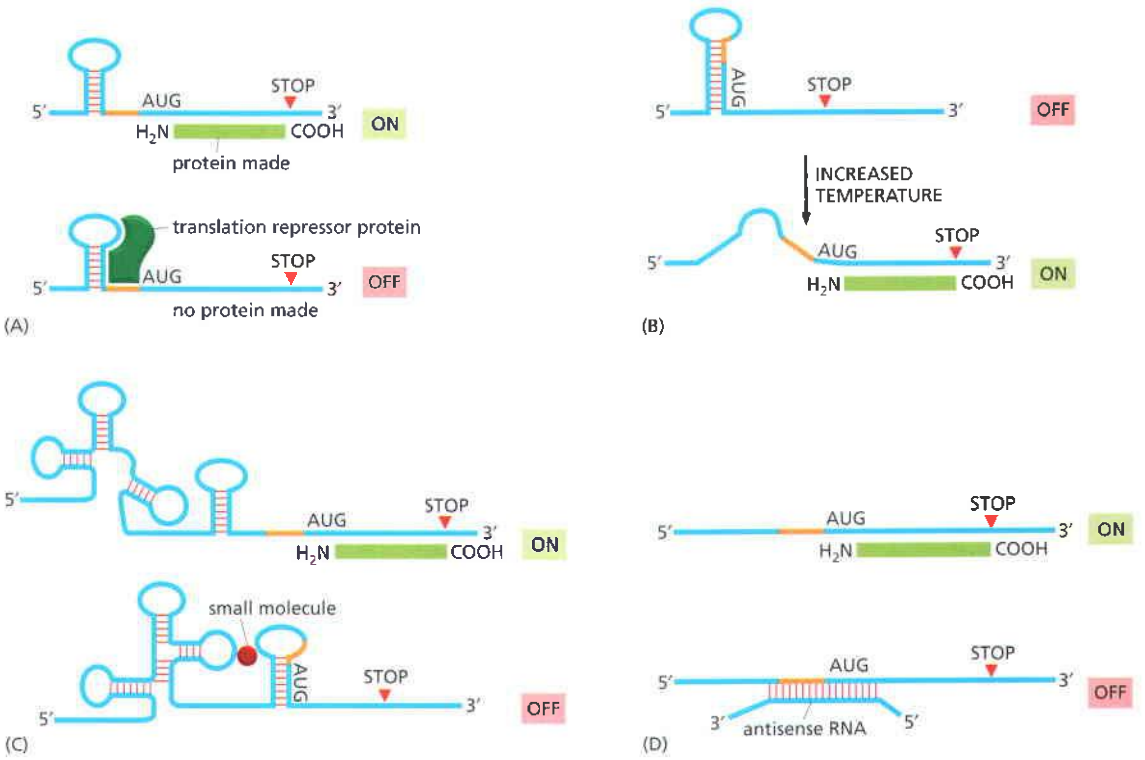


Figure 7–106 Mechanisms of translational control. Although these examples are from bacteria, many of the same principles operate in eucaryotes. (A) Sequence-specific RNA binding proteins repress translation of specific mRNAs by blocking access of the ribosome to the Shine–Dalgarno sequence (*orange*). For example, some ribosomal proteins repress translation of their own RNA. This mechanism comes into play only when the ribosomal proteins are produced in excess over ribosomal RNA and are therefore not incorporated into ribosomes, and it allows the cell to maintain correctly balanced quantities of the various components needed to form ribosomes. In these cases, the regulatory RNA sequence present on the mRNA often matches the RNA sequence that the protein recognizes during ribosome assembly. (B) An RNA "thermosensor" permits efficient translation initiation only at elevated temperatures in which the stem-loop structure has been melted. An example occurs in the human pathogen *Listeria monocytogenes*, in which the translation of its virulence genes increases at 37°C, the temperature of the host. (C) Binding of a small molecule to a riboswitch causes a structural rearrangement of the RNA, sequestering the Shine–Dalgarno sequence (*orange*) and blocking translation initiation. In many bacteria, *S*-adenosyl methionine acts in this manner to block production of the enzymes that synthesize it. (D) An "antisense" RNA produced elsewhere from the genome base-pairs with a specific mRNA, and blocks its translation. Many bacteria regulate expression of iron-storage proteins in this way. When iron is abundant, an antisense transcript is down-regulated, thereby allowing efficient translation of genes encoding the storage proteins. Antisense RNAs are used extensively by eucaryotic cells to regulate gene expression. The mechanism is somewhat different from that shown here and is discussed in detail later in this chapter.

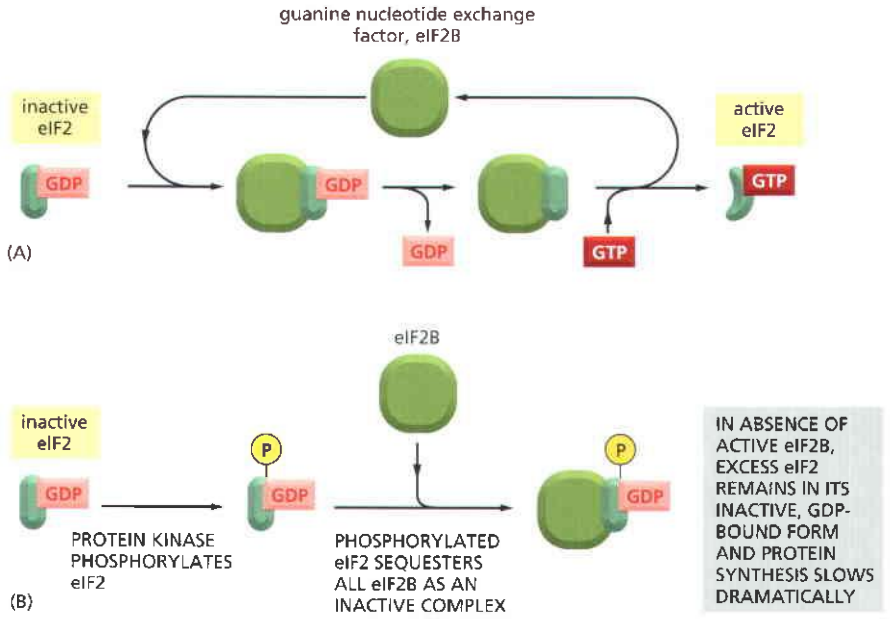


Figure 7–107 The eIF2 cycle. (A) The recycling of used eIF2 by a guanine nucleotide exchange factor (eIF2B). (B) eIF2 phosphorylation controls protein synthesis rates by tying up eIF2B.

AUG codon in the mRNA and skip to the second or third AUG codon instead. This phenomenon, known as "leaky scanning," is a strategy frequently used to produce two or more closely related proteins, differing only in their amino termini, from the same mRNA. Very importantly it allows some genes to produce the same protein with and without a signal sequence attached at its amino terminus so that the protein is directed to two different locations in the cell (for example, to both mitochondria and the cytosol). In some cases, the cell can regulate the relative abundance of the protein isoforms produced by leaky scanning; for example, a cell-type-specific increase in the abundance of the initiation factor eIF4F favors the use of the AUG closest to the 5′ end of the mRNA.

Another type of control found in eucaryotes uses one or more short open reading frames (nucleotide sequences free from stop codons) that lie between the 5′ end of the mRNA and the beginning of the gene. Open reading frames (ORFs) will be discussed more fully in Chapter 8: for present purposes an ORF can be considered a stretch of DNA that begins with a start codon (ATG) and ends with a stop codon, with no stop codons in between, and thus could in principle encode a polypeptide. Often, the amino acid sequences coded by these upstream open reading frames (uORFs) are not important; rather the uORFs serve a purely regulatory function. An uORF present on an mRNA molecule will generally decrease translation of the downstream gene by trapping a scanning ribosome initiation complex and causing the ribosome to translate the uORF and dissociate from the mRNA before it reaches the protein-coding sequences.

When the activity of a general translation factor (such as the eIF2 discussed above) is reduced, one might expect that the translation of all mRNAs would be reduced equally. Contrary to this expectation, however, the phosphorylation of eIF2 can have selective effects, even enhancing the translation of specific mRNAs that contain uORFs. This can enable yeast cells, for example, to adapt to starvation for specific nutrients by shutting down the synthesis of all proteins except those that are required for synthesis of the nutrients that are missing. The details of this mechanism have been worked out for a specific yeast mRNA that encodes a protein called Gcn4, a gene regulatory protein that is required for the activation of many genes that encode proteins that are important for amino acid synthesis.

The *Gcn4* mRNA contains four short uORFs, and these are responsible for selectively increasing the translation of *Gcn4* in response to the eIF2 phosphorylation provoked by amino acid starvation. The mechanism by which *Gcn4* translation is increased is complex. In outline, the small subunit of the ribosome moves along the mRNA, encountering each of the uORFs but directing translation of only a subset of them; if the last uORF is translated, as is the case in normal unstarved cells, the ribosomes dissociate at the end of the uORF, and translation of *Gcn4* is inefficient. The global decrease in eIF2 activity brought about

by nutrient starvation (see Figure 7–108) makes it more likely that a scanning small ribosomal subunit will move through the fourth uORF before it acquires a molecule of eIF2 (see Figure 6–72). Such a ribosomal subunit is free to initiate translation on the actual *Gcn4* sequences, and the increased amount of this gene regulatory protein that results leads to the production of a set of proteins that increase amino acid synthesis inside the cell.

Internal Ribosome Entry Sites Provide Opportunities for Translation Control

Although approximately 90% of eucaryotic mRNAs are translated beginning with the first AUG downstream from the 5′ cap, certain AUGs, as we saw in the previous section, can be skipped over during the scanning process. In this section, we discuss yet another way that cells can initiate translation at positions distant from the 5′ end of the mRNA, using a specialized type of RNA sequence called an **internal ribosome entry site** (**IRES**). An IRES can occur in many different places in an mRNA and, in some unusual cases, two distinct protein-coding sequences are carried in tandem on the same eucaryotic mRNA; translation of the first occurs by the usual scanning mechanism, and translation of the second occurs through an IRES. IRESs are typically several hundred nucleotides in length and fold into specific structures that bind many, but not all, of the same proteins that are used to initiate normal 5′ cap-dependent translation (**Figure 7–108**). In fact, different IRESs require different subsets of initiation factors. However, all of them bypass the need for a 5′ cap structure and the translation initiation factor that recognizes it, eIF4E.

Some viruses use IRESs as part of a strategy to get their own mRNA molecules translated while blocking normal 5′-cap-dependent translation of host mRNAs. On infection, these viruses produce a protease (encoded in the viral genome) that cleaves the host cell translation factor eIF4G, rendering it unable to bind to eIF4E, the cap-binding complex. This shuts down most of the host cell's translation and effectively diverts the translation machinery to the IRES sequences, which are present on many viral mRNAs. The truncated eIF4G remains competent to initiate translation at these internal sites and may even stimulate the translation of certain IRES-containing viral mRNAs. A selective activation of IRES-mediated translation can also occur on host cell mRNAs. For example, when mammalian cells enter the programmed cell death pathway (discussed in Chapter 18), eIF4G is cleaved, and a general decrease in translation ensues. However, some proteins critical for the control of cell death seem to be translated from IRES-containing mRNAs, allowing their continued synthesis. In this way, the IRES mechanism allows translation of selected mRNAs at a high rate despite a general decrease in the cell's overall capacity to initiate protein synthesis.

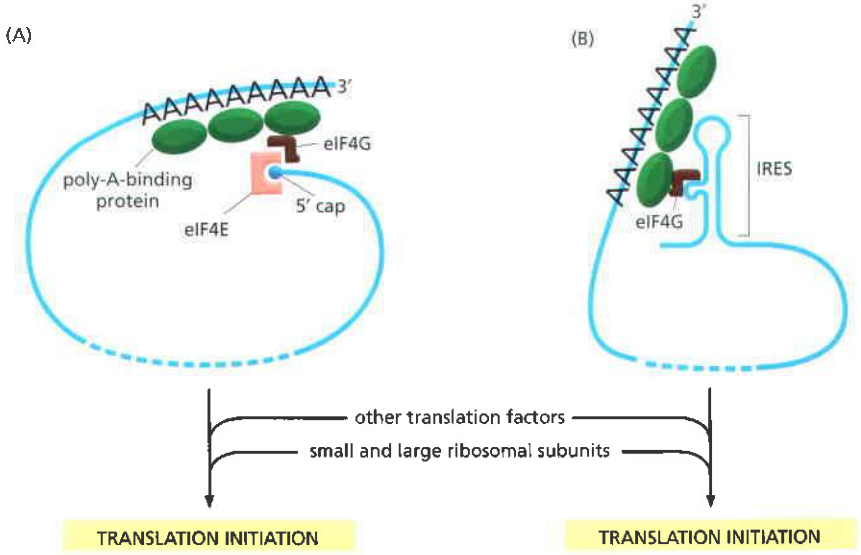


Figure 7–108 Two mechanisms of translation initiation. (A) The normal, cap-dependent mechanism requires a set of initiation factors whose assembly on the mRNA is stimulated by the presence of a 5′ cap and a poly-A tail (see also Figure 6–72). (B) The IRES-dependent mechanism seen mainly in viruses, requires only a subset of the normal translation initiating factors, and these assemble directly on the folded IRES. (Adapted from A. Sachs, *Cell* 101:243–245, 2000. With permission from Elsevier.)

Changes in mRNA Stability Can Regulate Gene Expression

Most mRNAs in a bacterial cell are very unstable, having half-lives of less than a couple of minutes. Exonucleases, which degrade in the 3′-to-5′ direction, are usually responsible for the rapid destruction of these mRNAs. Because its mRNAs are both rapidly synthesized and rapidly degraded, a bacterium can adapt quickly to environmental changes.

As a general rule, the mRNAs in eucaryotic cells are more stable. Some, such as that encoding β-globin, have half-lives of more than 10 hours, but most have considerably shorter half-lives, typically less than 30 minutes. The mRNAs that code for proteins such as growth factors and gene regulatory proteins, whose production rates need to change rapidly in cells, have especially short half-lives.

Two general mechanisms exist for destroying eucaryotic mRNAs. Both begin with the gradual shortening of the poly-A tail by an exonuclease, a process that starts as soon as the mRNA reaches the cytoplasm. In a broad sense, this poly-A shortening acts as a timer that counts down the lifetime of each mRNA. Once the poly-A tail is reduced to a critical length (about 25 nucleotides in humans), the two pathways diverge. In one, the 5′ cap is removed (a process called decapping) and the "exposed" mRNA is rapidly degraded from its 5′ end. In the other, the mRNA continues to be degraded from the 3′ end, through the poly-A tail into the coding sequences (**Figure 7–109**). Most eucaryotic mRNAs are degraded by both mechanisms.

Nearly all mRNAs are subject to these two types of decay, and the specific sequences of each mRNA determine how fast each step occurs and therefore how long each mRNA will persist in the cell and be able to produce protein. The 3′ UTR sequences are especially important in controlling mRNA lifetimes, and they often carry binding sites for specific proteins that increase or decrease the rate of poly-A shortening, decapping, or 3′-to-5′ degradation. The half-life of an mRNA is also affected by how efficiently it is translated. Poly-A shortening and decapping compete directly with the machinery that translates the mRNA; therefore any factors that affect the translation efficiency of an mRNA will tend to have the opposite effect on its degradation (**Figure 7–110**).

Although poly-A shortening controls the half-life of most eucaryotic mRNAs, some can be degraded by a specialized mechanism that bypasses this step altogether. In these cases, specific nucleases cleave the mRNA internally, effectively decapping one end and removing the poly-A tail from the other so that both halves are rapidly degraded. The mRNAs that are destroyed in this way carry specific nucleotide sequences, often in the 3′ UTRs, that serve as recognition sequences for these endonucleases. This strategy makes it especially simple to tightly regulate the stability of these mRNAs by blocking the endonuclease site in response to extracellular signals. For example, the addition of iron to cells decreases the stability of the mRNA that encodes the receptor protein that binds the iron-transporting protein transferrin, causing less of this receptor to be made. This effect is mediated by the iron-sensitive RNA-binding protein aconitase (which also controls ferritin mRNA translation). Aconitase can bind to the 3′ UTR of the transferrin receptor mRNA and increase receptor production by

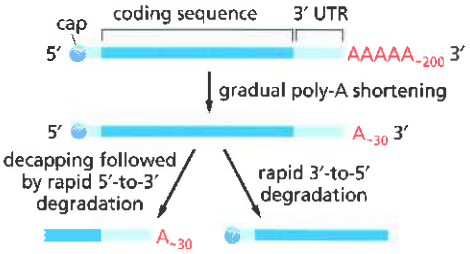


Figure 7–109 Two mechanisms of eucaryotic mRNA decay. A critical threshold of poly-A tail length induces 3′-to-5′ degradation, which may be triggered by the loss of the poly-A binding proteins (see Figure 6–40). As shown in Figure 7–110, the deadenylase associates with both the 3′ poly-A tail and the 5′ cap, and this arrangement may signal decapping after poly-A shortening. Although 5′ to 3′ and 3′ to 5′ degradation are shown here on separate RNA molecules, these two processes can occur together on the same molecule. (Adapted from C.A. Beelman and R. Parker, *Cell* 81:179–183, 1995. With permission from Elsevier.)

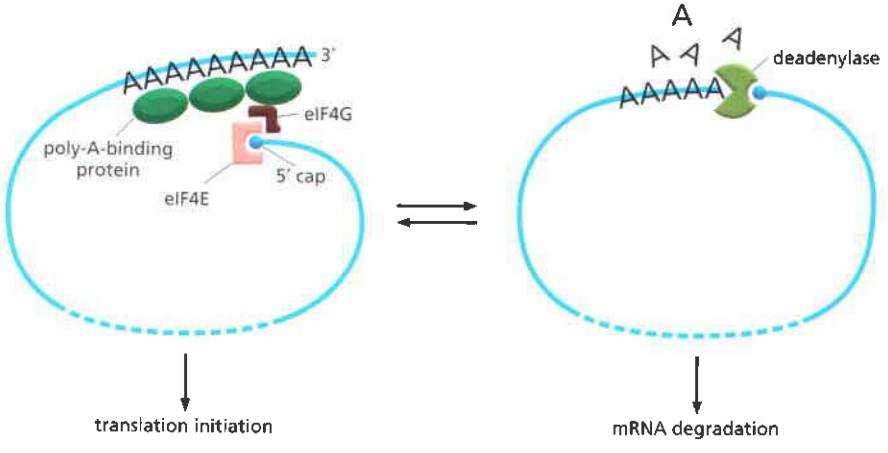


Figure 7–110 The competition between mRNA translation and mRNA decay. The same two features of an mRNA molecule, its 5′ cap and the 3′ poly-A tail, are used in both translation initiation and deadenylation-dependent mRNA decay (see Figure 7–109). The deadenylase that shortens the poly-A tail in the 3′-to-5′ direction associates with the 5′ cap. As described in Chapter 6 (see Figure 6–72), the translation initiation machinery also associates with both the 5′ cap and the poly-A tail. (Adapted from M. Gao et al., *Mol. Cell* 5:479–488, 2000. With permission from Elsevier.)

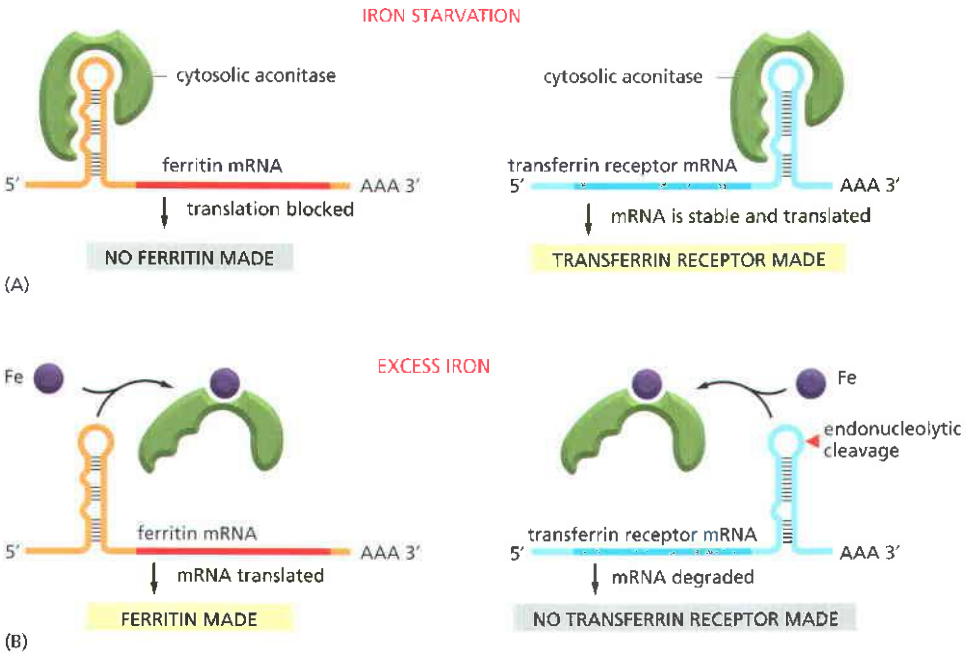


Figure 7–111 Two post-translational controls mediated by iron. (A) During iron starvation, the binding of aconitase to the 5′ UTR of the ferritin mRNA blocks translation initiation; its binding to the 3′ UTR of the transferrin receptor mRNA blocks an endonuclease cleavage site and thereby stabilizes the mRNA. (B) In response to an increase in iron concentration in the cytosol, a cell increases its synthesis of ferritin in order to bind the extra iron and decreases its synthesis of transferrin receptors in order to import less iron across the plasma membrane. Both responses are mediated by the same iron-responsive regulatory protein, aconitase, which recognizes common features in a stem-and-loop structure in the mRNAs encoding ferritin and transferrin receptor. Aconitase dissociates from the mRNA when it binds iron. But because the transferrin receptor and ferritin are regulated by different types of mechanisms, their levels respond oppositely to iron concentrations even though they are regulated by the same iron-responsive regulatory protein. (Adapted from M.W. Hentze et al., *Science* 238:1570–1573, 1987 and J.L. Casey et al., *Science* 240:924–928, 1988. With permission from AAAS.)

blocking endonucleolytic cleavage of the mRNA. On the addition of iron, aconitase is released from the mRNA, exposing the cleavage site and thereby decreasing the stability of the mRNA (Figure 7–111).

Cytoplasmic Poly-A Addition Can Regulate Translation

The initial polyadenylation of an RNA molecule (discussed in Chapter 6) occurs in the nucleus, apparently automatically for nearly all eucaryotic mRNA precursors. As we have just seen, the poly-A tails on most mRNAs gradually shorten in the cytosol, and the RNAs are eventually degraded. In some cases, however, the poly-A tails of specific mRNAs are lengthened in the cytosol, and this mechanism provides an additional form of translation regulation.

Maturing oocytes and eggs provide the most striking example. Many of the normal mRNA degradation pathways seem to be disabled in these giant cells, so that the cells can build up large stores of mRNAs in preparation for fertilization. Many mRNAs are stored in the cytoplasm with only 10 to 30 As at their 3′ end, and in this form they are not translated. At specific times during oocyte maturation and just after fertilization, when the cell requires the proteins encoded by these mRNAs, poly-A is added to selected mRNAs by a cytosolic poly-A polymerase, greatly stimulating their translation.

Small Noncoding RNA Transcripts Regulate Many Animal and Plant Genes

In the previous chapter, we introduced the central dogma, according to which the flow of genetic information proceeds from DNA through RNA to protein (Figure 6–2). But we have seen that RNA molecules perform many critical tasks in the cell besides serving as intermediate carriers of genetic information. A series of recent, striking discoveries has revealed that noncoding RNAs are far more prevalent than previously imagined and play previously unanticipated, but widespread, roles in regulating gene expression.

Of special importance to animals and plants is a type of short noncoding RNA called **microRNA** (miRNA). Humans, for example, express more than 400 different miRNAs, and these appear to regulate at least one-third of all human protein-coding genes. Once made, miRNAs base-pair with specific mRNAs and regulate their stability and their translation. The miRNA precursors are synthesized by RNA

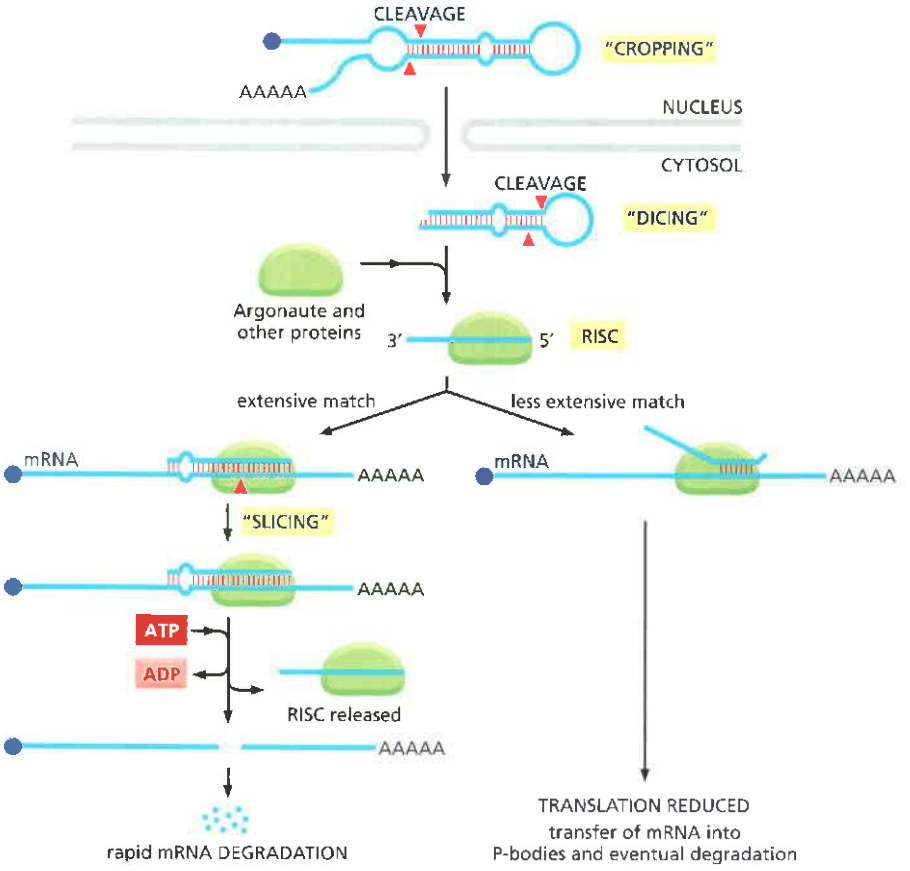


Figure 7–112 miRNA processing and mechanism of action. The precursor miRNA, through complementarity between one part of its sequence and another, forms a double-stranded structure. This is cropped while still in the nucleus, and then exported to the cytosol, where it is further cleaved by the Dicer enzyme to form the miRNA proper. Argonaute, in conjunction with other components of RISC, initially associates with both strands of the miRNA and cleaves and discards one of them. The other strand guides RISC to specific mRNAs through base pairing. If the RNA:RNA match is extensive, as is commonly seen in plants, Argonaute cleaves the target mRNA, causing its rapid degradation. In animals, the miRNA-mRNA match often does not extend beyond a short 7-nucleotide "seed" region near the 5′ end of the miRNA. This less extensive base pairing leads to inhibition of translation, mRNA destabilization, and transfer of the mRNA to P-bodies, where it is eventually degraded.

polymerase II and are capped and polyadenylated. They then undergo a special type of processing, after which the miRNA is assembled with a set of proteins to form an *RNA-induced silencing complex* or *RISC*. Once formed, the RISC seeks out its target mRNAs by searching for complementary nucleotide sequences (**Figure 7–112**). This search is greatly facilitated by the Argonaute protein, a component of RISC, which displays the 5′ region of the miRNA so that it is optimally positioned for base-pairing to another RNA molecule (**Figure 7–113**). In animals, the extent of base-pairing is typically seven nucleotide pairs, and it usually takes place in the 3′ UTR of the target mRNA.

Once an mRNA has been bound by an miRNA, several outcomes are possible. If the base-pairing is extensive, the mRNA is cleaved by the Argonaute protein, effectively removing its poly-A tail and exposing it to exonucleases (see Figure 7–109). Following cleavage of the mRNA, RISC (with its associated miRNA) is released, and it can seek out additional mRNAs. Thus, a single miRNA can act catalytically to destroy many complementary mRNAs. The miRNAs can be thought of as guide sequences that bring destructive nucleases into contact with specific mRNAs.

If the base-pairing between the miRNA and the mRNA is less extensive, Argonaute does not slice the mRNA; rather, translation of the mRNA is repressed and the mRNA is destabilized. This effect is associated with shortening of the poly-A tail and the movement of the mRNA to cytosolic structures called *processing bodies (P-bodies)*. Here the mRNAs are sequestered from ribosomes and eventually decapped and degraded. P-bodies are dynamic structures composed of large assemblies of mRNAs and RNA-degrading enzymes, and they are believed to be the sites in the cell where the final destruction of most mRNAs, even those not controlled by miRNAs, takes place (**Figure 7–114**).

Several features make miRNAs especially useful regulators of gene expression. First, a single miRNA can regulate a whole set of different mRNAs so long as the mRNAs carry a common sequence in their UTRs. This situation is common in humans, where some miRNAs control hundreds of different mRNAs. Second, regulation by miRNAs can be combinatorial. When the base-pairing between the miRNA and mRNA fails to trigger cleavage, additional miRNAs binding to the

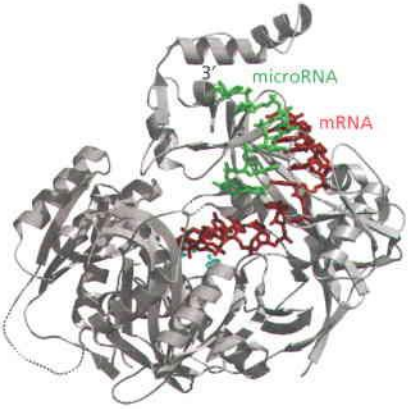


Figure 7–113 Structure of Argonaute protein bound to a perfectly base-paired miRNA and mRNA. (Adapted from N.H. Tolia and L. Joshua-Tor, *Nat. Chem. Biol.* 3:36–43, 2007. With permission from Macmillan Publishers Ltd.)

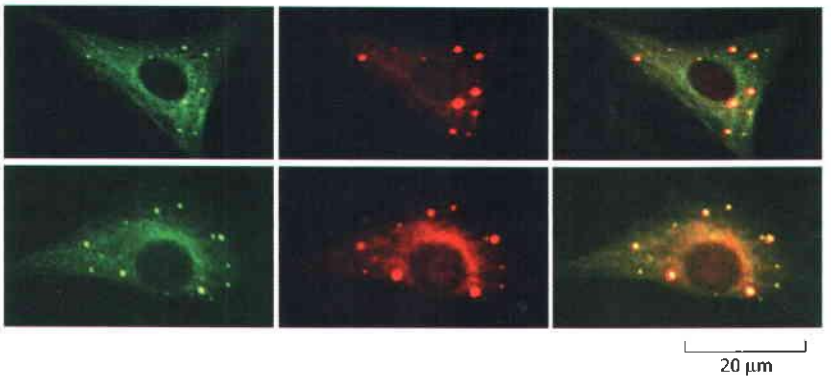


Figure 7–114 Visualization of P-bodies. Human cells were stained with antibodies to a component of the mRNA decapping enzyme Dcp1a *(left panels)* and to the Argonaute protein *(middle panels)*. The merged image *(right panels)* shows that the two proteins co-localize to foci in the cytoplasm called P-bodies. (Adapted from J. Liu et al., *Nat. Cell Biol.* 7:643-644, 2005. With permission from Macmillan Publishers Ltd.)

same mRNA lead to further reductions in its translation. As discussed earlier in this chapter for gene regulatory proteins, combinatorial control greatly expands the possibilities available to the cell by linking gene expression to a combination of different regulators rather than a single regulator. Third, an miRNA occupies relatively little space in the genome when compared with a protein. Indeed, their small size is one reason that miRNAs were discovered only recently. Although we are only beginning to understand the full impact of miRNAs, it is clear that they represent a very important part of the cell's equipment for regulating the expression of its genes.

RNA Interference Is a Cell Defense Mechanism

Many of the proteins that participate in the miRNA regulatory mechanisms just described also serve a second function as a defense mechanism: they orchestrate degradation of foreign RNA molecules, specifically those that occur in double-stranded form. Termed **RNA interference** (**RNAi**), this mechanism is found in a wide variety of organisms, including single-celled fungi, plants, and worms, suggesting that it is evolutionarily ancient. Many transposable elements and viruses produce double-stranded RNA, at least transiently, in their life cycles, and RNAi helps to keep these potentially dangerous invaders in check. As we shall see, RNAi has also provided scientists with a powerful experimental technique to turn off the expression of individual genes.

The presence of double-stranded RNA in the cell triggers RNAi by attracting a protein complex containing *Dicer*, the same nuclease that processes miRNAs (see Figure 7–112). This protein complex cleaves the double-stranded RNA into small (approximately 23- nucleotide-pair) fragments called **small interfering RNAs** (**siRNA**). These double-stranded siRNAs are then bound by Argonaute and other components of the RISC, as we saw above for miRNAs, and one strand of the duplex RNA is cleaved by Argonaute and discarded. The single-stranded siRNA molecule that remains directs RISC back to complementary RNA molecules produced by the virus or transposable element; because the match is exact, Argonaute cleaves these molecules, leading to their rapid destruction (**Figure 7–115**).

Each time RISC cleaves a new RNA molecule, it is released; thus as we saw for miRNAs, a single RNA molecule can act catalytically to destroy many complementary RNAs. Some organisms employ an additional mechanism that amplifies the RNAi response even further. In these organisms, RNA-dependent RNA polymerases can convert the products of siRNA-mediated cleavage into more double-stranded RNA. This amplification ensures that, once initiated, RNA interference can continue even after all the initiating double-stranded RNA has been degraded or diluted out. For example, it permits progeny cells to continue carrying out RNA interference that was provoked in the parent cells.

In some organisms, the RNA interference activity can be spread by the transfer of RNA fragments from cell to cell. This is particularly important in plants (whose cells are linked by fine connecting channels, as discussed in Chapter 19), because it allows an entire plant to become resistant to an RNA virus after only a few of its cells have been infected. In a broad sense, the RNAi

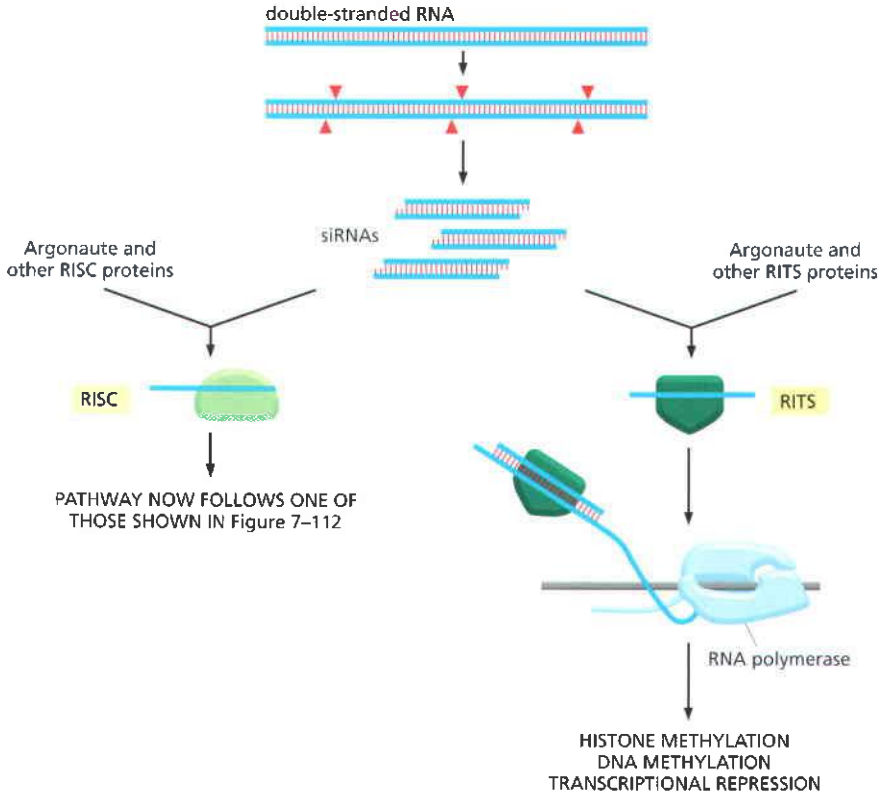


Figure 7–115 siRNA-mediated heterochromatin formation. In many organisms, double-stranded RNA can trigger both the destruction of complementary mRNAs *(left)* and transcriptional silencing *(right)*. The change in chromatin structure induced by the bound RITS (RNA-induced transcriptional silencing) complex resembles that in Figure 7–81.

response resembles certain aspects of animal immune systems; in both, an invading organism elicits a customized response, and—through amplification of the "attack" molecules—the host becomes systemically protected.

RNA Interference Can Direct Heterochromatin Formation

The RNA interference pathway just described does not necessarily stop with the destruction of target RNA molecules. In some cases, the RNA interference machinery can selectively shut off synthesis of the target RNAs. For this remarkable mechanism to occur, the short siRNAs produced by the Dicer protein are assembled with a group of proteins (including Argonaute) to form the RITS (RNA-induced transcriptional silencing) complex. Using single-stranded siRNA as a guide sequence, this complex binds complementary RNA transcripts as they emerge from a transcribing RNA polymerase II (see Figure 7–115). Positioned on the genome in this manner, the RITS complex attracts proteins that covalently modify nearby histones and eventually directs the formation and spread of heterochromatin to prevent further transcription initiation. In some cases, the RITS complex also induces the methylation of DNA, which, as we have seen, can repress gene expression even further. Because heterochromatin and DNA methylation can be self-propagating, an initial RNA interference signal can continue to silence gene expression long after all the siRNA molecules have dissipated.

RNAi-directed heterochromatin formation is an important cell defense mechanism that limits the accumulation of transposable elements in the genome by maintaining them in a transcriptionally silent form. However, this same mechanism is also used in many normal processes in the cell. For example, in many organisms, the RNA interference machinery maintains the heterochromatin formed around centromeres. Centromeric DNA sequences are transcribed in both directions, producing complementary RNA transcripts that can base-pair to form double-stranded RNA. This double-stranded RNA triggers the RNA interference pathway and stimulates heterochromatin formation at centromeres. This heterochromatin, in turn, is necessary for the centromeres to segregate chromosomes accurately during mitosis (see Figure 4–50).

RNA Interference Has Become a Powerful Experimental Tool

Although it probably arose initially as a defense mechanism, RNA interference has become thoroughly integrated into many aspects of normal cell biology, ranging from the control of gene expression to the structure of chromosomes. It has also been developed by scientists into a powerful experimental tool that allows almost any gene to be inactivated by evoking an RNAi response to it. This technique, carried out in cultured cells, and in some cases, whole animals and plants, has revolutionized genetic approaches in cell and molecular biology. We shall discuss it in more detail in the following chapter (see pp. 571–572). RNAi also has great potential in treating human disease. Since many human disorders result from the misexpression of genes, the ability to turn these genes off by experimentally introducing complementary siRNA molecules holds great medical promise. Remarkably, the mechanism of RNA interference was discovered only recently, and we are still being surprised by its mechanistic details and by its broad biological implications.

Summary

Many steps in the pathway from RNA to protein are regulated by cells in order to control gene expression. Most genes are regulated at multiple levels, in addition to being controlled at the initiation stage of transcription. The regulatory mechanisms include (1) attenuation of the RNA transcript by its premature termination, (2) alternative RNA splice-site selection, (3) control of 3′-end formation by cleavage and poly-A addition, (4) RNA editing, (5) control of transport from the nucleus to the cytosol, (6) localization of mRNAs to particular parts of the cell, (7) control of translation initiation, and (8) regulated mRNA degradation. Most of these control processes require the recognition of specific sequences or structures in the RNA molecule being regulated, a task performed by either regulatory proteins or regulatory RNA molecules. A particularly widespread form of post-transcriptional control is RNA interference, where guide RNAs base-pair with mRNAs. RNA interference can cause mRNAs to be either destroyed or translationally repressed. It can also cause specific genes to be packaged into heterochromatin.

PROBLEMS

Which statements are true? Explain why or why not.

7–1 In terms of its biochemical function, the helix-loop-helix motif is more closely related to the leucine zipper motif than it is to the helix-turn-helix motif.

7–2 Reversible genetic rearrangements are a common way of regulating gene expression in procaryotes and mammalian cells.

7–3 CG islands are thought to have arisen during evolution because they were associated with portions of the genome that remained active, hence unmethylated, in the germline.

Discuss the following problems.

7–4 A small portion of a two-dimensional display of proteins from human brain is shown in **Figure Q7–1**. These proteins were separated on the basis of size in one dimension and electrical charge (isoelectric point) in the other. Not all protein spots on such displays are products of different genes; some represent modified forms of a protein that migrate to different positions. Pick out a couple of sets of spots that could represent proteins that differ by the number of phosphates they carry. Explain the basis for your selection.

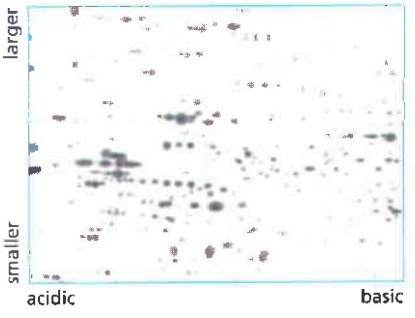


Figure Q7–1 Two-dimensional separation of proteins from the human brain (Problem 7–4). The proteins were displayed using two-dimensional gel electrophoresis. Only a small portion of the protein spectrum is shown. (Courtesy of Tim Myers and Leigh Anderson, Large Scale Biology Corporation.)

7–5 DNA microarray analysis of the patterns of mRNA abundance in different human cell types shows that the level of expression of almost every active gene is different. The patterns of mRNA abundance are so characteristic of cell type that they can be used to determine the tissue of origin of cancer cells, even though the cells may have metastasized to different parts of the body. By definition, however, cancer cells are different from their noncancerous precursor cells. How do you suppose then that patterns of mRNA expression might be used to determine the tissue source of a human cancer?

7–6 The nucleus of a eucaryotic cell is much larger than a bacterium, and it contains much more DNA. As a consequence, a DNA-binding protein in a eucaryotic cell must be able to select its specific binding site from among many more unrelated sequences than does a DNA-binding protein in a bacterium. Does this present any special problems for eucaryotic gene regulation?

Consider the following situation. Assume that the eucaryotic nucleus and the bacterial cell each have a single copy of the same DNA-binding site. In addition, assume that the nucleus is 500 times the volume of the bacterium, and has 500 times as much DNA. If concentration of the gene regulatory protein that binds the site were the same in the nucleus and in the bacterium, would the regulatory protein find its binding site equally as well in the eucaryotic nucleus as it does in the bacterium? Explain your answer.

7–7 DNA-binding proteins often find their specific sites much faster than would be anticipated by simple three-dimensional diffusion. The Lac repressor, for example, associates with the *Lac* operator—its DNA-binding site—more than 100 times faster than expected from this model. Clearly, the repressor must find the operator by mechanisms that reduce the dimensionality or volume of the search in order to hasten target acquisition.

Several techniques have been used to investigate this problem. One of the most elegant used strongly fluorescent RNA polymerase molecules that could be followed individually. An array of DNA molecules was aligned in parallel and anchored to a glass slide. Fluorescent RNA polymerase molecules were then allowed to flow across them at an oblique angle (**Figure Q7–2**A). Traces of individual RNA polymerases showed that about half flowed in the same direction as the bulk and about half deviated from the bulk flow in a characteristic manner (Figure Q7–2B). If the RNA polymerase molecules were first incubated with short DNA fragments containing a strong promoter, all the traces followed the bulk flow.

A. Offer an explanation for why some RNA polymerase molecules deviated from the bulk flow as shown in Figure Q7–2B. Why did incubation with short DNA fragments containing a strong promoter eliminate traces that deviated from the bulk flow?

B. Do these results suggest an explanation for how site-specific DNA-binding molecules manage to find their sites faster than expected by diffusion?

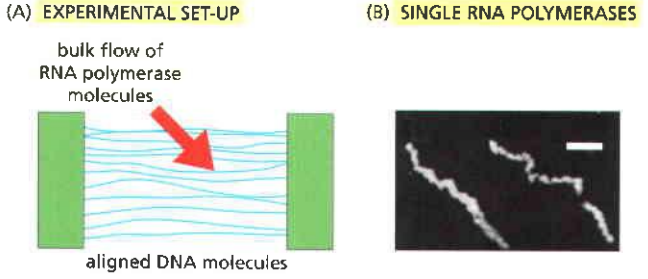


Figure Q7–2 Interactions of individual RNA polymerase molecules with DNA (Problem 7–7). (A) Experimental set-up. DNA molecules are aligned and anchored to a glass slide, and highly fluorescent RNA polymerase molecules are allowed to flow across them. (B) Traces of two individual RNA polymerase molecules. The one on the *left* has traveled with the bulk flow, and the one on the *right* has deviated from it. The scale bar is 10 μm. (B, reprinted from H. Kabata et al., *Science* 262:1561–1563, 1993. With permission from AAAS.

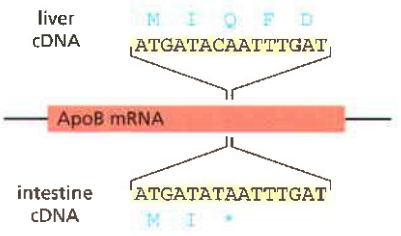


Figure Q7–3 Location of the sequence differences in cDNA clones from ApoB RNA isolated from liver and intestine (Problem 7–9). The encoded amino acid sequences, in the one-letter code, are shown aligned with the cDNA sequences.

C. Based on your explanation, would you expect a site-specific DNA-binding molecule to find its target site faster in a population of short DNA molecules or in a population of long DNA molecules? Assume that the concentration of target sites is identical and that there is one target site per DNA molecule.

7–8 Most people who are completely blind have circadian rhythms that are 'free-running;' that is, their rhythms are not synchronized to environmental time cues and they oscillate on a cycle of about 24.5 hours. Why do you suppose the circadian clocks of blind people are not entrained to the same 24-hour clock as the majority of the population? Can you guess what symptoms might be associated with a free-running circadian clock? Do you suppose that blind people have trouble sleeping?

7–9 In humans, two closely related forms of apolipoprotein B (ApoB) are found in blood as constituents of the plasma lipoproteins. ApoB48 (molecular mass, 48 kilodaltons) is synthesized by the intestine and is a key component of chylomicrons, the large lipoprotein particles responsible for delivery of dietary triglycerides to adipose tissue for storage. ApoB100 (molecular mass, 100,000 kilodaltons) is synthesized in the liver for formation of the much smaller, very low-density lipoprotein particles used in the distribution of triglycerides to meet energy needs. A classic set of studies defined the surprising relationship between these two proteins.

Sequences of cloned cDNA copies of the mRNAs from these two tissues revealed a single difference: cDNAs from intestinal cells had a T, as part of a stop codon, at a point where the cDNAs from liver cells had a C, as part of a glutamine codon (**Figure Q7–3**). To verify the differences in the mRNAs and to search for corresponding differences in the genome, RNA and DNA were isolated from intestinal and liver cells and then subjected to PCR amplification, using oligonucleotides that flanked the region of interest. The amplified DNA segments from the four samples were tested for the presence of the T or C by hybridization to oligonucleotides containing either the liver cDNA sequence (oligo-Q) or the intestinal cDNA sequence (oligo-STOP). The results are shown in **Table Q7–1**.

Are the two forms of ApoB produced by transcriptional control from two different genes, by a processing control of the RNA transcript from a single gene, or by differential cleavage of the protein product from a single gene? Explain your reasoning.

Table Q7–1 Hybridization of specific oligonucleotides to the amplified segments from liver and intestine RNA and DNA (Problem 7–9).

	RNA		DNA	
	LIVER	INTESTINE	LIVER	INTESTINE
Oligo-Q	+	–	+	+
Oligo-STOP	–	+	–	–

Hybridization is indicated by +; its absence by –.

REFERENCES

General

Brown TA (2002) Genomes 2, 2nd ed. New York: Wiley-Liss.
Epigenetics (2004) *Cold Spring Harb Symp Quant Biol* 69.
Hartwell L, Hood L, Goldberg ML et al (2006) Genetics: from Genes to Genomes, 3rd ed. Boston: McGraw Hill.
Lodish H, Berk A, Kaiser CL et al (2007) Molecular Cell Biology, 6th ed. New York: WH Freeman.
McKnight SL & Yamamoto KR (eds) (1993) Transcriptional Regulation. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.
Mechanisms of Transcription (1998) *Cold Spring Harb Symp Quant Biol* 63.
Ptashne M & Gann A (2002) Genes and Signals. Cold Spring Harbor Laboratory Press, Cold Spring Harbor.
Regulatory RNAs (2006) *Cold Spring Harb Symp Quant Biol* 71.
Watson JD, Baker TA, Bell SP et al (2003) Molecular Biology of the Gene, 5th ed. Menlo Park, CA: Benjamin Cummings.

An Overview of Gene Control

Campbell KH, McWhir J, Ritchie WA & Wilmut I (1996) Sheep cloned by nuclear transfer from a cultured cell line. *Nature* 380:64–66.
Davidson EH (2006) The Regulatory Genome: Gene Regulatory Networks in Development and Evolution. Burlington, MA: Elsevier.
Gurdon JB (1992) The generation of diversity and pattern in animal development. *Cell* 68:185–199.
Levine M & Tjian R (2003) Transcription regulation and animal diversity. *Nature* 424:147–151.
Ross DT, Scherf U, Eisen MB et al (2000) Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genet* 24:227–235.

DNA-binding Motifs in Gene Regulatory Proteins

Gehring WJ, Affolter M & Burglin T (1994) Homeodomain proteins. *Annu Rev Biochem* 63:487–526.
Harbison CT, Gordon DB, Lee TI et al (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature* 431:99–104.
Luscombe NM, Austin SE, Berman HM, et al (2000) An overview of the structures of protein–DNA complexes. *Gen Biol* 1:reviews 001.1–001.37.
McKnight SL (1991) Molecular zippers in gene regulation. *Sci Am* 264:54–64.
Pabo CO & Sauer RT (1992) Transcription factors: structural families and principles of DNA recognition. *Annu Rev Biochem* 61:1053–1095.
Rhodes D & Klug A (1993) Zinc fingers. *Sci Am* 268:56–65.
Seeman NC, Rosenberg JM & Rich A (1976) Sequence-specific recognition of double helical nucleic acids by proteins. *Proc Natl Acad Sci USA* 73:804–808.

How Genetic Switches Work

Becker PB & Hörz W (2002) ATP-dependent nucleosome remodeling. *Annu Rev Biochem* 71:247–273.
Beckwith J (1987) The operon: an historical account. In *Escherichia coli* and *Salmonella typhimurium*: Cellular and Molecular Biology (Neidhart FC, Ingraham JL, Low KB et al eds), vol 2, pp 1439–1443. Washington, DC: ASM Press.
Gaszner M & Felsenfeld G (2006) Insulators: exploiting transcriptional and epigenetic mechanisms. *Nature Rev Genet* 7:703–713.
Gilbert W and Müller-Hill B (1967) The lac operator is DNA. *Proc Natl Acad Sci USA* 58:2415.
Green MR (2005) Eukaryotic transcription activation: right on target. *Mol Cell* 18:399–402.
Jacob F & Monod J (1961) Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol* 3:318–356.
Lawson CL, Swigon D, Murakami KS et al (2004) Catabolite activator protein: DNA binding and transcription activation. *Curr Opin Struct Biol* 14:10–20.
Millar CB, & Grunstein M (2006) Genome-wide patterns of histone modifications in yeast. *Nature Rev Mol Cell Biol* 7:657–666.
Narlikar GJ, Fan HY & Kingston RE (2002) Cooperation between complexes that regulate chromatin structure and transcription. *Cell* 108:475–487.
Oehler S, Eismann ER, Krämer H et al (1990) The three operators of the lac operon cooperate in repression. *EMBO J* 9:973–979.
Ptashne M (2004) A Genetic Switch: Phage and Lambda Revisited, 3rd ed. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.
Ptashne M (1967) Specific binding of the lambda phage repressor to lambda DNA. *Nature* 214:232–234.
St Johnston D & Nusslein-Volhard C (1992) The origin of pattern and polarity in the *Drosophila* embryo. *Cell* 68:201–219.
Strahl BD & Allis CD (2000) The language of covalent histone modifications. *Nature* 403:41–45.

The Molecular Genetic Mechanisms that Create Specialized Cell Types

Alon, U (2006) An Introduction to Systems Biology: Design Principles of Biological Circuits (Chapman & Hall/Crc Mathematical and Computational Biology Series) TF Chapman.
Bell-Pedersen D, Cassone VM, Earnest DJ et al (2005) Circadian rhythms from multiple oscillators: lessons from diverse organisms. *Nature Rev Genet* 6:544–556.
Bernstein BE, Meissner A & Lander ES (2007) The mammalian epigenome. *Cell* 128:669–681.
Herskowitz I (1989) A regulatory hierarchy for cell specialization in yeast. *Nature* 342:749–757.
Klose RJ & Bird AP (2006) Genomic DNA methylation: the mark and its mediators. *Trends Biochem Sci* 31:89–97.
Meyer BJ (2000) Sex in the worm: counting and compensating X-chromosome dose. *Trends Genet* 16:247–253.
Surani MA (2001) Reprogramming of genome function through epigenetic inheritance. *Nature* 414:122–128.
Tapscott SJ (2005) The circuitry of a master switch: MyoD and the regulation of skeletal muscle gene transcription. *Development* 132:2685–2695.

Post-transcriptional Controls

Bass BL (2002) RNA editing by adenosine deaminases that act on RNA. *Annu Rev Biochem* 71:817–846.
Blencowe BJ (2006) Alternative splicing: new insights from global analyses. *Cell* 126:37–47.
Brennecke J, Stark A, Russell RB et al (2005) Principles of microRNA-target recognition. *PLoS Biology* 3.
Fire A, Xu S, Montgomery MK et al (1998) Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature* 391:806–811.
Frankel AD & Young JAT (1998) HIV-1: fifteen proteins and an RNA. *Annu Rev Biochem* 67:1–25.
Gottesman S (2004) The small RNA regulators of *Escherichia coli*: roles and mechanisms. *Annu Rev Microbiol* 58:303–328.
Mello CC & Conte D (2004) Revealing the world of RNA interference. *Nature* 431:338–342.
Parker R & Sheth U (2007) P bodies and the control of mRNA translation and degradation. *Mol Cell* 25:635–646.
Stuart KD, Schnaufer A, Ernst NL et al (2005) Complex management: RNA editing in trypanosomes. *Trends Biochem Sci* 30:97–105.
Tolia NH & Joshua-Tor L (2007) Slicer and the argonautes. *Nature Chem Biol* 3:36–43.
Tomari Y & Zamore PD (2005) Perspective: machines for RNAi. *Genes Dev* 19:517–529.
Valencia-Sanchez MA, Liu J, Hannon GJ et al (2006) Control of translation and mRNA degradation by miRNAs and siRNAs. *Genes Dev* 20:515–524.
Verdel A & Moazed D (2005) RNAi-directed assembly of heterochromatin in fission yeast. *FEBS Letters* 579:5872–5878.
Wilhelm JE & Smibert, CA (2005) Mechanisms of translational regulation in *Drosophila*. *Biol Cell* 97:235–252.
Winkler WC & Breaker RR (2005) Regulation of bacterial gene expression by riboswitches. *Annu Rev Microbiol* 59:487–517.