Garland Science

Molecular Biology of THE CELL

Sixth Edition

ALBERTS JOHNSON LEWIS MORGAN RAFF ROBERTS WALTER

PART II

BASIC GENETIC MECHANISMS

CHAPTER 4

DNA, Chromosomes, and Genomes

IN THIS CHAPTER

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. The 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 exert precise control, directing multicellular development as well as the daily life of every cell? What kind of instructions does the genetic information contain? And 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 largely of instructions for making proteins. Proteins are phenomenally versatile macromolecules that perform most cell functions. As we saw in Chapter 3, they serve as building blocks for cell structures and form the enzymes that catalyze most of the cell's chemical reactions. They also 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 cells and organisms are determined to a great extent by the proteins that they are able to make.

Painstaking observations of cells and embryos in the late nineteenth century had led to the recognition that the hereditary information is carried on *chromosomes*—threadlike structures in the nucleus of a eukaryotic cell that become visible by light microscopy as the cell begins to divide (**Figure 4–1**). Later, when biochemical analysis became possible, chromosomes were found to consist of deoxyribonucleic acid (DNA) and protein, with both being present in roughly the same amounts. For many decades, the DNA was thought to be merely a structural

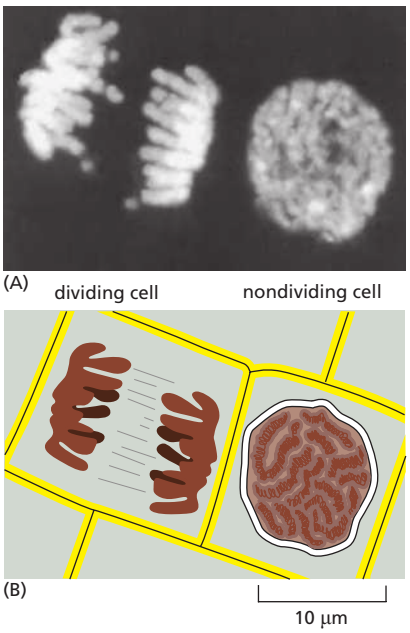


Figure 4–1 Chromosomes in cells. (A) Two adjacent plant cells photographed through a light microscope. The DNA has been stained with a fluorescent dye (DAPI) that binds to it. The DNA is present in chromosomes, which become visible as distinct structures in the light microscope only when they become 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 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.)

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 of inheritance in bacteria (Figure 4–2). But still, 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 puzzle was suddenly solved in 1953, when James Watson and Francis Crick derived the mechanism from their model of DNA structure. As outlined in Chapter 1, the determination of 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 by this breakthrough discovery.

We begin this chapter 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. And it must also allow access to chromosomal DNA, both for the enzymes that repair DNA damage and for the specialized proteins that direct the expression of its many genes.

In the past two decades, there has been a revolution in our ability to determine the exact order of subunits in DNA molecules. As a result, we now know the

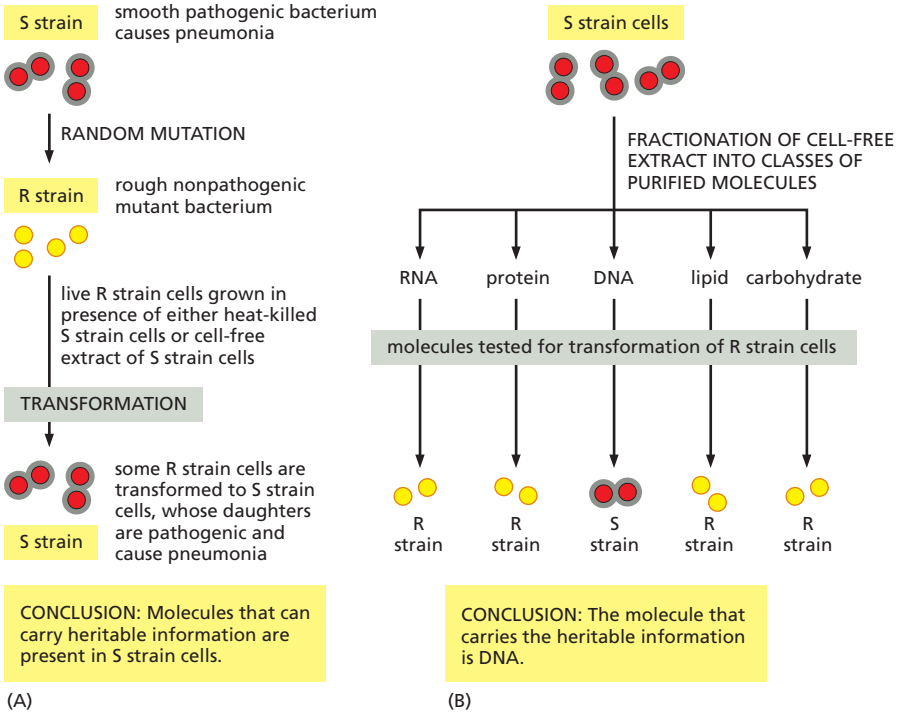


Figure 4–2 The first experimental demonstration that DNA is the genetic material. These experiments, carried out in the 1920s (A) and 1940s (B), showed that adding purified DNA to a bacterium changed the bacterium's properties and that this change was faithfully passed on to subsequent generations. Two 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 some 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 active substance as DNA.

sequence of the 3.2 billion nucleotide pairs that provide the information for producing a human adult from a fertilized egg, as well as having the DNA sequences for thousands of other organisms. Detailed analyses of these sequences are providing 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, and expresses the genetic information carried in its DNA. In the next chapter (Chapter 5), we shall discuss the mechanisms by which the cell accurately replicates and repairs DNA; we also describe how DNA sequences can be 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 is manufactured only at the proper time and place in the life of a cell.

THE STRUCTURE AND FUNCTION OF DNA

Biologists in the 1940s had difficulty in conceiving how DNA could be the genetic material. The molecule seemed too simple: a long polymer composed of only four types of nucleotide 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 provided one of the major clues that led to the Watson–Crick model for DNA structure that, as soon as it was proposed in 1953, made DNA's potential for replication and information storage apparent.

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*. The chains run antiparallel to each other, and *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. 100–101), 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 nucleotide subunit, each polynucleotide chain in DNA is analogous to a sugar-phosphate necklace (the backbone), from which hang the four types of beads (the bases A, C, G, and T). These same symbols (A, C, G, and T) are commonly used to denote either the four bases or the four entire nucleotides—that is, the bases with their attached sugar and phosphate groups.

The way in which the nucleotides 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*, names derived from the orientation of the deoxyribose sugar. With respect to DNA's

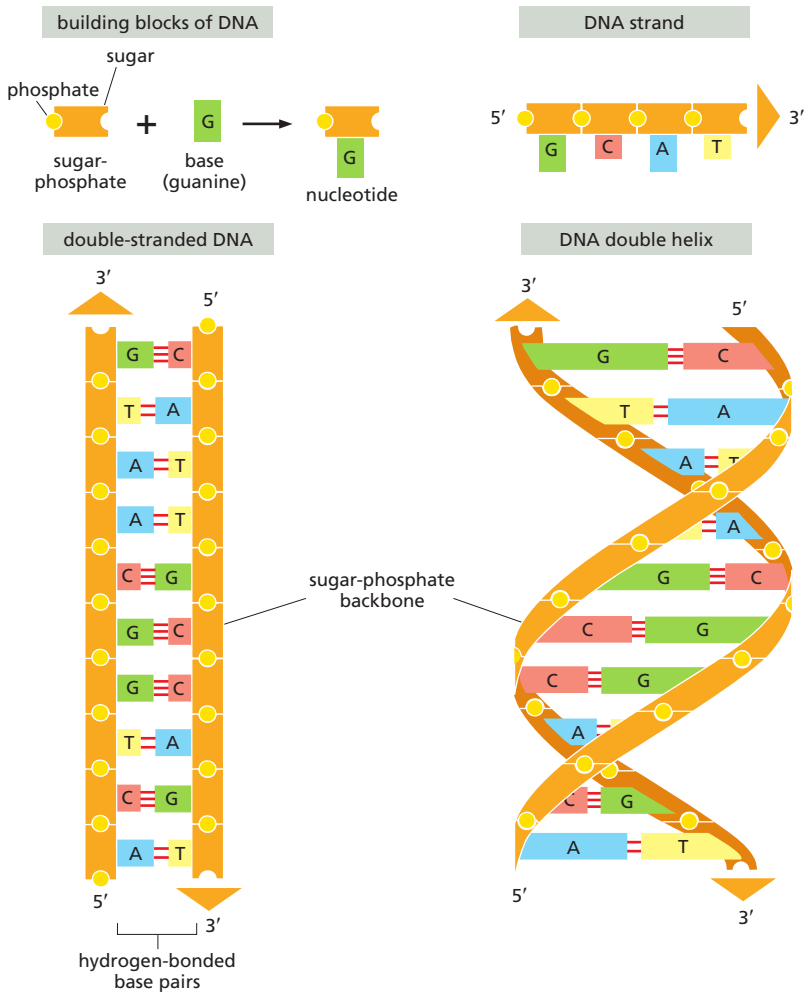


Figure 4–3 DNA and its building blocks. DNA is made of four types of nucleotides, which are linked covalently into a polynucleotide chain (a DNA strand) with a sugar-phosphate backbone from which the bases (A, C, G, and T) extend. A DNA molecule is composed of two antiparallel DNA strands held together by hydrogen bonds between the paired bases. The *arrowheads* at the ends of the DNA strands indicate the polarities of the two strands. 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 and Movie 4.1.

information-carrying capacity, the chain of nucleotides in a DNA strand, being both directional and linear, can be read in much the same way as the letters on this page.

The three-dimensional structure of DNA—the DNA **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. 100–101) is paired with a single-ring base (a pyrimidine): A always pairs with T, and G with C (**Figure 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 a constant distance apart along the DNA molecule. To maximize the efficiency of base-pair packing, the two sugar-phosphate backbones wind around each other to form a right-handed 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 DNA's structure and 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.

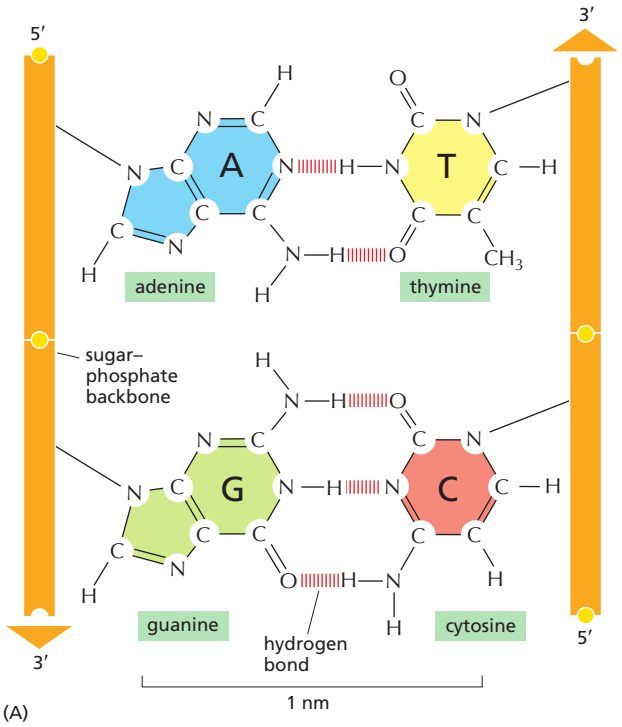


Figure 4–4 Complementary base pairs in the DNA double helix. The shapes and chemical structures of the bases allow hydrogen bonds to form efficiently only between A and T and between G and C, because atoms that are able to form hydrogen bonds (see Panel 2–3, pp. 94–95) can then be brought close together 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.

The Structure of DNA Provides a Mechanism for Heredity

The discovery of the structure of DNA immediately suggested answers to the two most fundamental questions about heredity. First, how could the information to specify an organism be carried in a chemical form? And second, how could this information be duplicated and copied from generation to generation?

The answer to the first question came from the realization that DNA is a linear polymer of four different kinds of monomer, strung out in a defined sequence like the letters of a document written in an alphabetic script.

The answer to the second question came from the double-stranded nature of the structure: 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

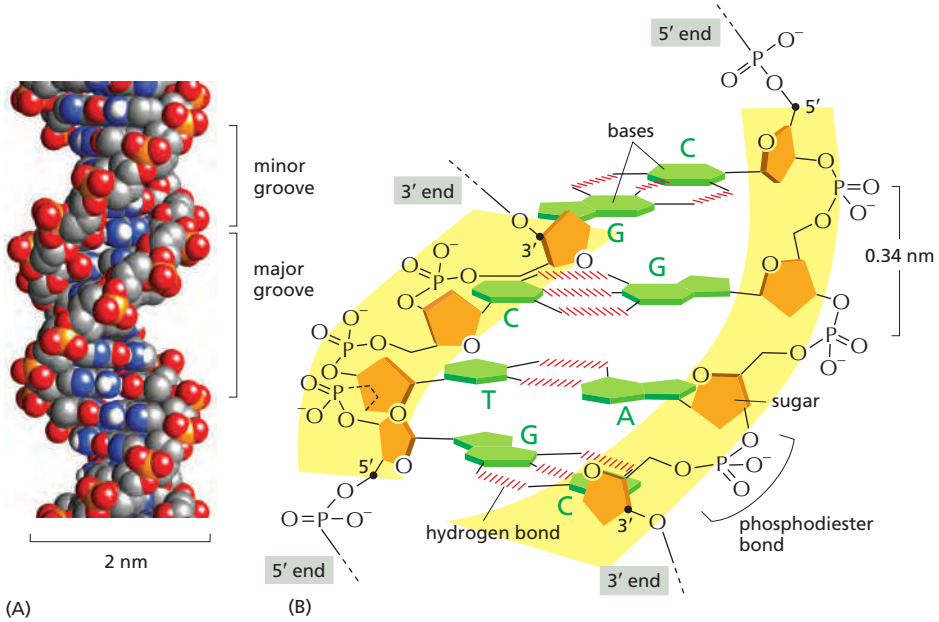


Figure 4–5 The DNA double helix. (A) A space-filling model of 1.5 turns of the DNA double helix. Each turn of DNA is made up of 10.4 nucleotide pairs, and the center-to-center distance between adjacent nucleotide pairs is 0.34 nm. The coiling of the two strands around each other creates two grooves in the double helix: the wider groove is called the major groove, and the smaller the minor groove, as indicated. (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.

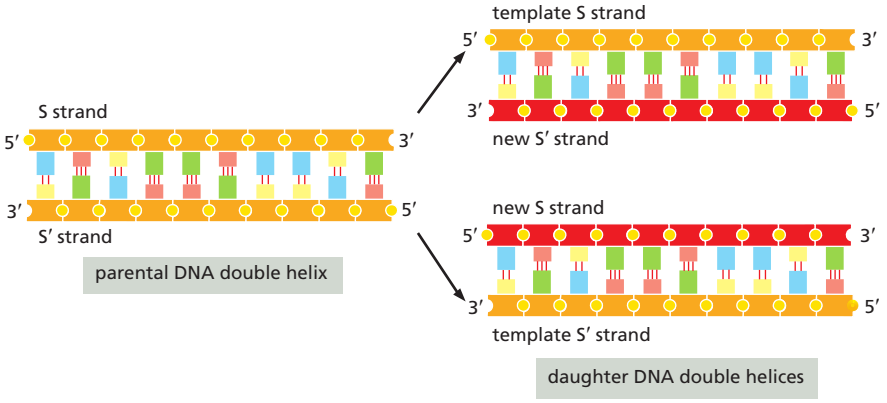


Figure 4–6 DNA as a template for its own duplication. Because the nucleotide A successfully pairs only with T, and G pairs with C, each strand of DNA can act as a template to specify the sequence of nucleotides in its complementary strand. In this way, double-helical DNA can be copied precisely, with each parental DNA helix producing two identical daughter DNA helices.

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 4–6**). 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. We shall describe the elegant machinery that the cell uses to perform this task in Chapter 5.

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. If genes are made of DNA, the DNA must therefore somehow encode proteins (**Figure 4–7**). 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 at all obvious from the DNA structure, and it took over a decade after the discovery of the double helix before it was worked out. In Chapter 6, we will describe this code in detail in the course of elaborating the process of *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 store of information in an organism's DNA is called its **genome**, and it specifies all the RNA molecules and proteins 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. The nucleotide sequence of a very small human gene, written out in the four-letter nucleotide alphabet, occupies a quarter of a page of text (**Figure 4–8**), 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 includes roughly 21,000 protein-coding genes, which (through alternative splicing; see p. 415) give rise to a much greater number of distinct proteins.

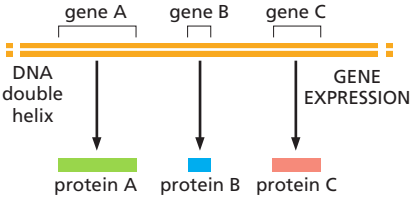


Figure 4–7 The relationship between genetic information carried in DNA and proteins. (Discussed in Chapter 1.)

In Eukaryotes, DNA Is Enclosed in a Cell Nucleus

As described in Chapter 1, nearly all the DNA in a eukaryotic 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

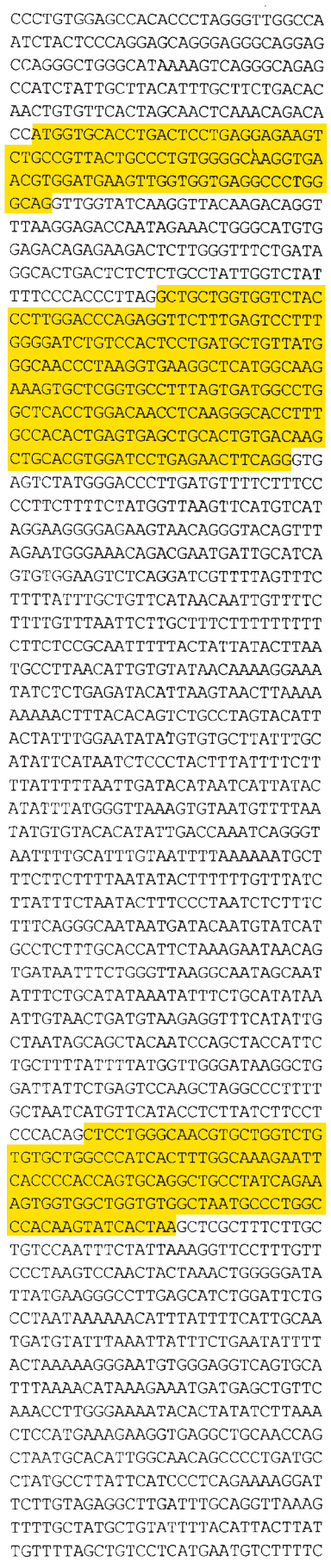


Figure 4–7 The nucleotide sequence of the human β-globin gene. By convention, a nucleotide sequence is written from its 5′ end to its 3′ end, and it should be read from left to right in successive lines down the page as though it were normal English text. This gene carries the information for the amino acid sequence of one of the two types of subunits of the hemoglobin molecule; a different gene, the α-globin gene, carries the information for the other. (Hemoglobin, the protein that carries oxygen in the blood, 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.

bilayer membranes (**Figure 4–9**). These membranes are punctured at intervals by large nuclear pores, through which molecules move between the nucleus and the cytosol. The nuclear envelope is directly connected to the extensive system of intracellular membranes called 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*—a thin feltlike mesh 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 chapters, it also keeps nuclear and cytosolic enzymes separate, a feature that is crucial for the proper functioning of eukaryotic cells.

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 antiparallel 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 RNA molecules and proteins that the organism will ever synthesize and is said to comprise its genome. In eukaryotes, DNA is contained in the cell nucleus, a large membrane-bound compartment.

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 RNA molecules and proteins that make up an organism—including information about when, in what types of cells, and in what quantity each RNA molecule and protein is to be made. The nuclear DNA of eukaryotes is 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 as a separate entity 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 the DNA and fold it, generating a series of organized coils and loops that provide increasingly higher levels of organization, and prevent the DNA from becoming an unmanageable tangle. Amazingly, although the DNA is very tightly compacted, it nevertheless remains accessible to the many enzymes in the cell that replicate it, repair it, and use its genes to produce RNA molecules and proteins.

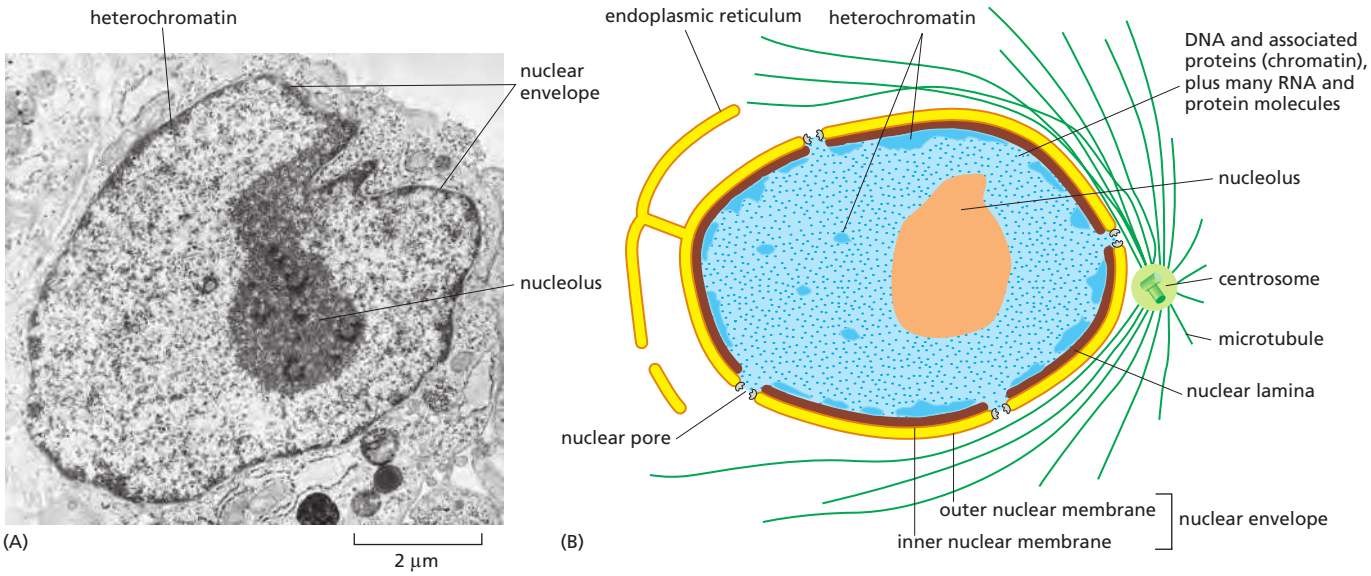


Figure 4–9 A cross-sectional view of a typical cell nucleus. (A) Electron micrograph of a thin section through the nucleus of a human fibroblast. (B) Schematic drawing, showing that the nuclear envelope consists of two membranes, the outer one being continuous with the endoplasmic reticulum (ER) membrane (see also Figure 12–7). 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 sheetlike network of intermediate filaments *(brown)* inside the nucleus forms the nuclear lamina *(brown),* providing mechanical support for the nuclear envelope (for details, see Chapter 12). The dark-staining heterochromatin contains specially condensed regions of DNA that will be discussed later. (A, courtesy of E.G. Jordan and J. McGovern.)

Eukaryotic DNA Is Packaged into a Set of Chromosomes

Each **chromosome** in a eukaryotic cell consists of a single, enormously long linear DNA molecule along with the proteins that fold and pack the fine DNA thread into a more compact structure. In addition to the proteins involved in packaging, chromosomes are also associated with many other proteins (as well as numerous RNA molecules). These are required for the processes of gene expression, DNA replication, and DNA repair. The complex of DNA and tightly bound protein is called *chromatin* (from the Greek *chroma*, "color," because of its staining properties).

Bacteria lack a special nuclear compartment, and they generally carry their genes on a single DNA molecule, which is often circular (see Figure 1–24). This DNA is also associated with proteins that package and condense it, but they are different from the proteins that perform these functions in eukaryotes. Although the bacterial DNA with its attendant proteins is often called the bacterial "chromosome," it does not have the same structure as eukaryotic chromosomes, and less is known about how the bacterial DNA is packaged. Therefore, our discussion of chromosome structure will focus almost entirely on eukaryotic chromosomes.

With the exception of the gametes (eggs and sperm) and a few highly specialized cell types that cannot multiply and either lack DNA altogether (for example, red blood cells) or have replicated their DNA without completing cell division (for example, megakaryocytes), each human cell nucleus 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). These human chromosomes can be readily distinguished by "painting" each one a different color using a technique based on *DNA hybridization* (**Figure 4–10**). In this method (described in detail in Chapter 8), a short strand of nucleic acid tagged with a fluorescent dye serves as a "probe" that picks out its complementary DNA sequence, lighting up the target chromosome at any site where it binds. Chromosome painting is most

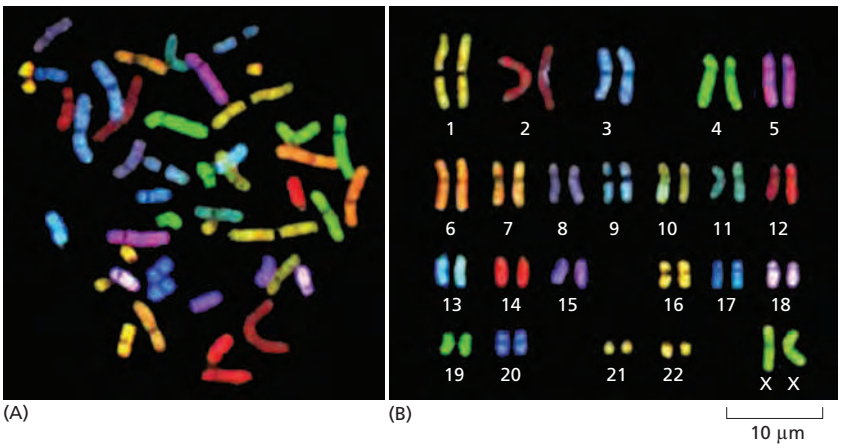


Figure 4–10 The complete set of human chromosomes. These chromosomes, from a female, were isolated from a cell undergoing nuclear division (mitosis) and are therefore highly compacted. Each chromosome has been "painted" a different color to permit its unambiguous identification under the fluorescence microscope, using a technique called "spectral karyotyping." Chromosome painting can be performed by exposing the chromosomes to a large collection of DNA molecules whose sequence matches known DNA sequences from the human genome. The set of sequences matching each chromosome is coupled to a different combination of fluorescent dyes. 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, each chromosome becomes labeled with a different combination of dyes. For such experiments, the chromosomes are subjected to treatments that separate the two strands of double-helical DNA in a way that permits base-pairing with the single-stranded labeled DNA, but keeps the overall 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. (Adapted from N. McNeil and T. Ried, *Expert Rev. Mol. Med.* 2:1–14, 2000. With permission from Cambridge University Press.)

frequently 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 is to stain them with dyes that reveal a striking and reproducible pattern of bands along each mitotic chromosome (Figure 4–11). These banding patterns presumably reflect variations in chromatin structure, but their basis is not well understood. Nevertheless, the pattern of bands on each type of chromosome is unique, and it provided the initial means to identify and number each human chromosome reliably.

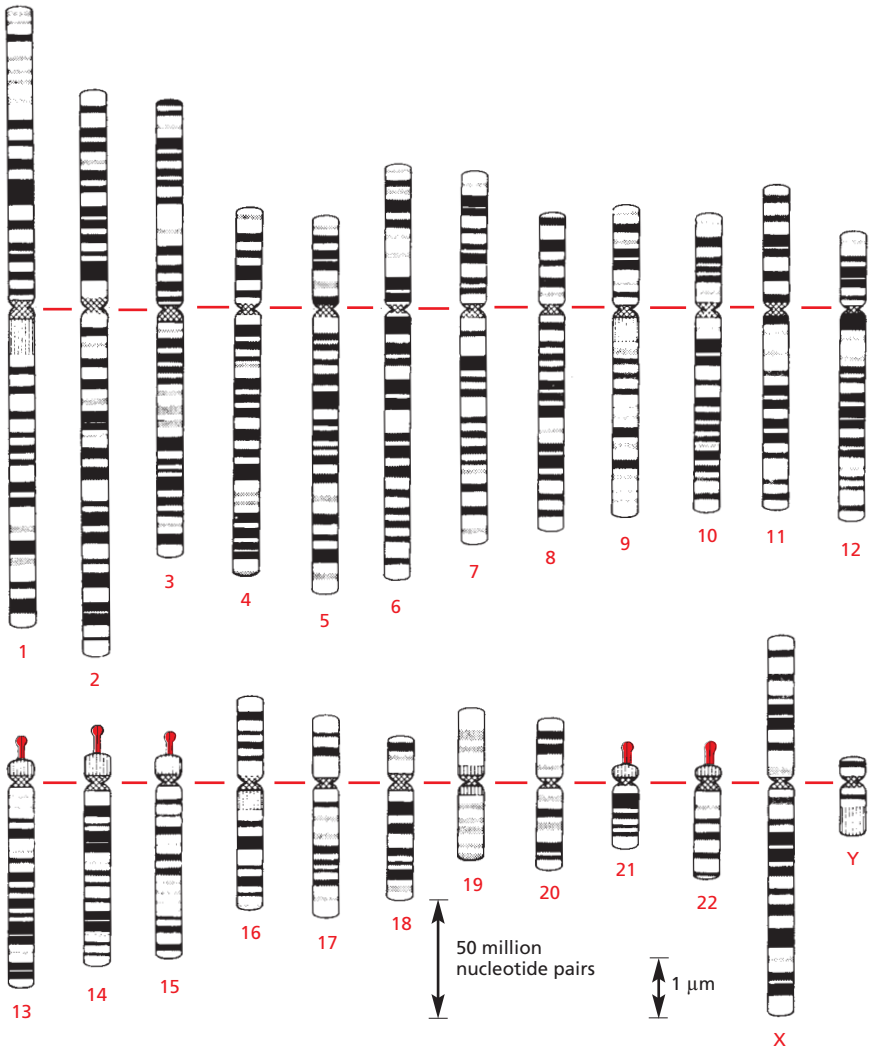


Figure 4–11 The banding patterns of human chromosomes. Chromosomes 1–22 are numbered in approximate order of size. A typical human cell contains two of each of these chromosomes, plus two sex chromosomes—two X chromosomes in a female, one X and one Y chromosome in a male. The chromosomes used to make 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 4–19), 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 banding patterns are obtained by staining chromosomes with Giemsa stain, and they can be observed under the light microscope. (Adapted from U. Francke, *Cytogenet. Cell Genet.* 31:24–32, 1981. With permission from the author.)

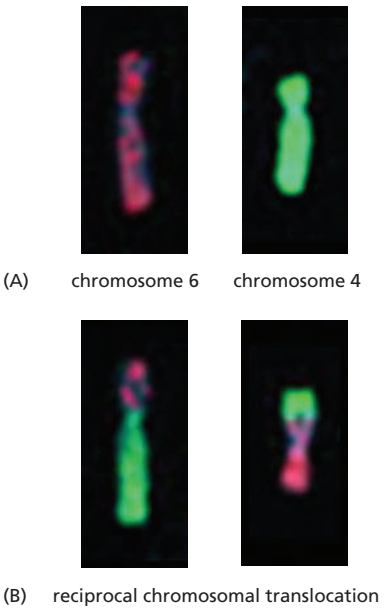


Figure 4–12 Aberrant human chromosomes. (A) Two normal human chromosomes, 4 and 6. (B) In an individual carrying a balanced *chromosomal translocation*, the DNA double helix in one chromosome has crossed over with the DNA double helix in the other chromosome due to an abnormal recombination event. The chromosome painting technique used on the chromosomes in each of the sets allows the identification of even short pieces of chromosomes that have become translocated, a frequent event in cancer cells. (Courtesy of Zhenya Tang and the NIGMS Human Genetic Cell Repository at the Coriell Institute for Medical Research: GM21880.)

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 either by changes in the banding patterns or—with greater sensitivity—by changes in the pattern of chromosome painting (**Figure 4–12**). Cytogeneticists use these alterations to detect inherited chromosome abnormalities and to reveal the chromosome rearrangements that occur in cancer cells as they progress to malignancy (discussed in Chapter 20).

Chromosomes Contain Long Strings of Genes

Chromosomes carry genes—the functional units of heredity. A gene is often defined as a segment of DNA that contains the instructions for making a particular protein (or a set of closely related proteins), but this definition is too narrow. Genes that code for protein are indeed the majority, and most of the genes with clear-cut mutant phenotypes fall under this heading. In addition, however, there are many "RNA genes"—segments of DNA that generate a functionally significant RNA molecule, instead of a protein, as their final product. We shall say more about the RNA genes and their products later.

As might be expected, some correlation exists between the complexity of an organism and the number of genes in its genome (see Table 1–2, p. 29). For example, some simple bacteria have only 500 genes, compared to about 30,000 for humans. Bacteria, archaea, and some single-celled eukaryotes, such as yeast, have concise genomes, consisting of little more than strings of closely packed genes. However, the genomes of multicellular plants and animals, as well as many other eukaryotes, contain, in addition to genes, a large quantity of interspersed DNA whose function is poorly understood (**Figure 4–13**). Some of this additional DNA is crucial for the proper control of gene expression, and this may in part explain why there is so much of it in multicellular organisms, whose genes have to be switched on and off according to complicated rules during development (discussed in Chapters 7 and 21).

Differences in the amount of DNA interspersed between genes, far more than differences in numbers of genes, account for the astonishing variations in genome size that we see when we compare one species with another (see Figure 1–32). For example, the human genome is 200 times larger than that of the yeast *Saccharomyces 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 noncoding DNA, the genomes of closely related 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

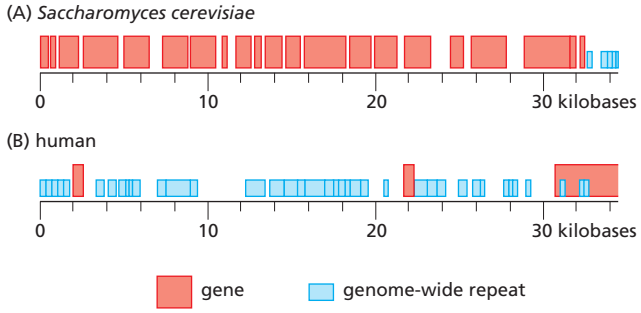
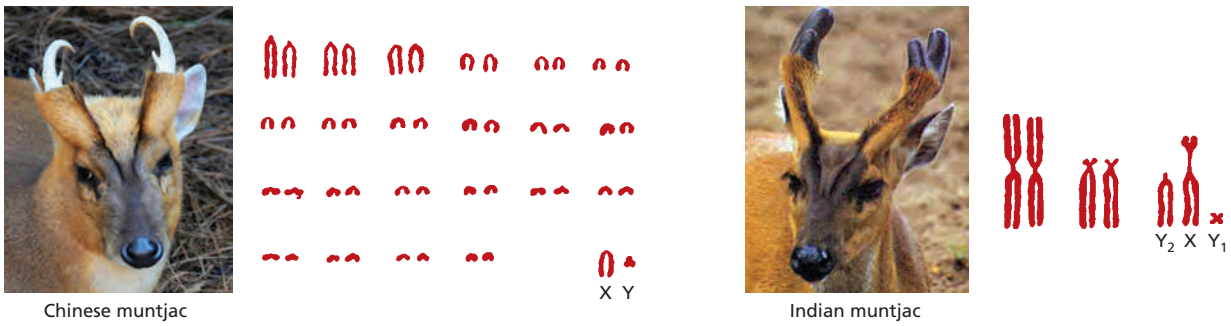


Figure 4–13 The arrangement of genes in the genome of *S. cerevisiae* compared to humans. (A) *S. cerevisiae* is a budding yeast widely used for brewing and baking. The genome of this single-celled eukaryote is distributed over 16 chromosomes. A small region of one chromosome has been arbitrarily selected to show its high density of genes. (B) A region of the human genome of equal length to the yeast segment in (A). The human genes are much less densely packed and the amount of interspersed DNA sequence is far greater. Not shown in this sample of human DNA is the fact that most human genes are much longer than yeast genes (see Figure 4–15).



DNA may do, it seems clear that it is not a great handicap for a eukaryotic cell to carry a large amount of it.

How the genome is divided into chromosomes also differs from one eukaryotic species to the next. For example, while the cells of humans have 46 chromosomes, those of some small deer have only 6, while those of the common 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, complexity of the organism, 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 poorly understood selection pressures over long evolutionary times.

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. (Chinese muntjac photo courtesy of Deborah Carreno, Natural Wonders Photography.)

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

With the publication of the full DNA sequence of the human genome in 2004, it became possible to see in detail how the genes are arranged along each of our chromosomes (**Figure 4–15**). It will be many decades before the information contained in the human genome sequence is fully analyzed, but it has already stimulated new experiments that have had major effects on the content of every chapter in this book.

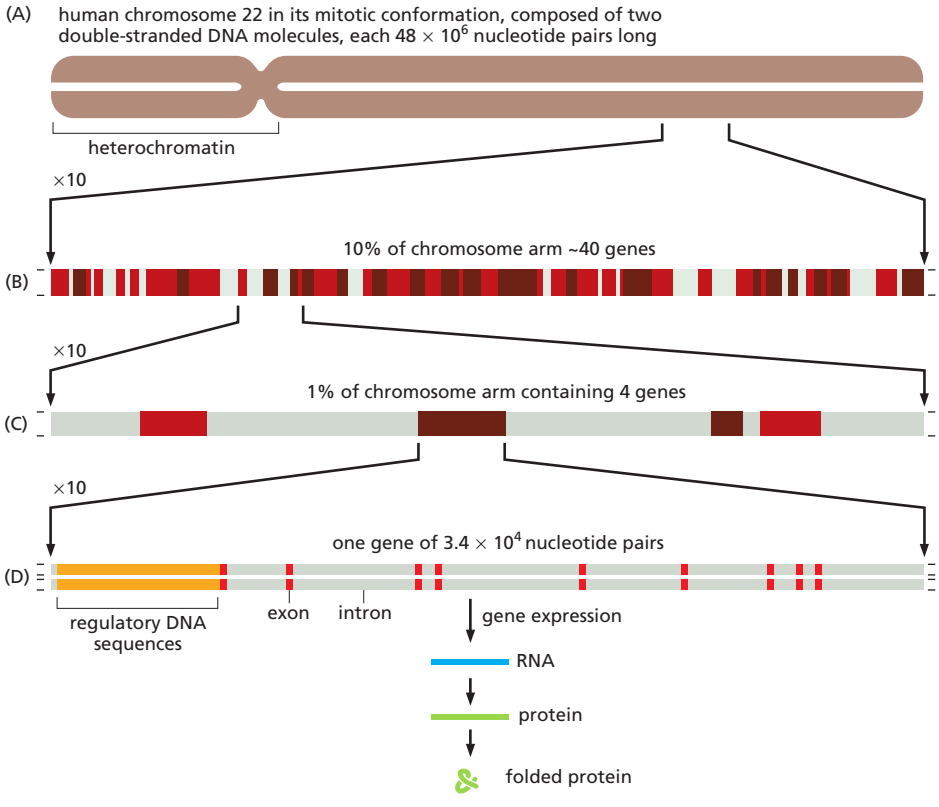


Figure 4–15 The organization of genes on a human chromosome. (A) Chromosome 22, one of the smallest human chromosomes, contains 48 × 10^6 nucleotide pairs and makes up approximately 1.5% of the human genome. Most of the left arm of chromosome 22 consists of short repeated sequences of DNA that are packaged in a particularly compact form of chromatin (heterochromatin) 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) showing four 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 different 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.)

TABLE 4–1 Some Vital Statistics for the Human Genome

Human genome	
DNA length	3.2×10^9 nucleotide pairs*
Number of genes coding for proteins	Approximately 21,000
Largest gene coding for protein	2.4×10^6 nucleotide pairs
Mean size for protein-coding genes	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 noncoding RNA genes	Approximately 9000**
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-number repetitive elements	Approximately 50%

* The sequence of 2.85 billion nucleotides is known precisely (error rate of only about 1 in 100,000 nucleotides). The remaining DNA primarily consists of short sequences that are tandemly repeated many times over, with repeat numbers differing from one individual to the next. These highly repetitive blocks are hard to sequence accurately.
** This number is only a very rough estimate.
*** A pseudogene is a DNA sequence closely resembling that of a functional gene, but containing numerous mutations that prevent its proper expression or function. Most pseudogenes arise from the duplication of a functional gene followed by the accumulation of damaging mutations in one copy.
**** These conserved functional regions include DNA encoding 5′ and 3′ UTRs (untranslated regions of mRNA), DNA specifying structural and functional RNAs, and DNA with conserved protein-binding sites.

The first striking feature of the human genome is how little of it (only a few percent) codes for proteins (Table 4–1 and Figure 4–16). It is also notable that nearly half of the chromosomal DNA is made up of mobile pieces of DNA that have gradually inserted themselves in the chromosomes over evolutionary time, multiplying like parasites in the genome (see Figure 4–62). We discuss these *transposable elements* in detail in later chapters.

A second notable feature of the human genome is the large average gene size—about 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 sequence 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.

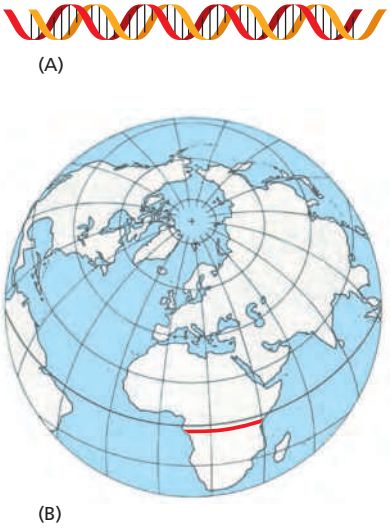


Figure 4–16 Scale of the human genome. If drawn with a 1 mm space between each nucleotide pair, as in (A), the human genome would extend 3200 km (approximately 2000 miles), far enough to stretch across the center of Africa, the site of our human origins (*red line* in B). At this scale, there would be, on average, a protein-coding gene every 150 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.

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 much more compressed in organisms with concise genomes. We discuss how regulatory DNA sequences work in Chapter 7.

Research in the last decade has surprised biologists with the discovery that, in addition to 21,000 protein-coding genes, the human genome contains many thousands of genes that encode RNA molecules that do not produce proteins, but instead have a variety of other important functions. What is thus far known about these molecules will be presented in Chapters 6 and 7. Last, but not least, the nucleotide sequence of the human genome has revealed that the archive of information needed to produce a human seems to be in an alarming state of chaos. 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." We shall discuss how this is thought to have come about in the final sections of this chapter entitled "How Genomes Evolve."

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

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 4–17**, and it is discussed in detail in Chapter 17. Briefly, during a long *interphase*, genes are expressed and chromosomes are replicated, with the two replicas remaining together as a pair of *sister chromatids*. Throughout this time, the chromosomes are extended and much of their chromatin exists as long threads in the nucleus so that individual chromosomes cannot be easily distinguished. It is only during a much briefer period of mitosis that each chromosome condenses so that its two sister chromatids can be separated and distributed to the two daughter nuclei. The highly condensed chromosomes in a dividing cell are known as *mitotic chromosomes* (**Figure 4–18**). 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.

Each 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

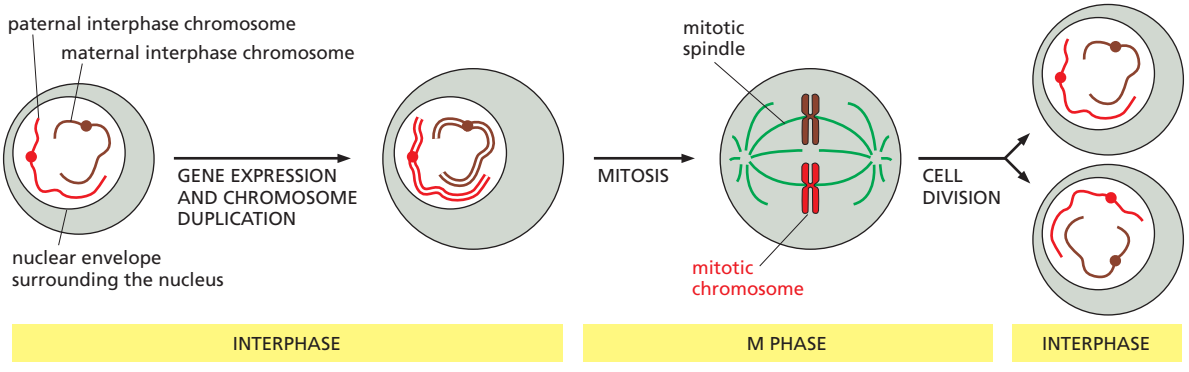


Figure 4–17 A simplified view of the eukaryotic cell cycle. During interphase, the cell is actively expressing its genes and is therefore synthesizing proteins. Also, during interphase and before cell division, the DNA is replicated and each chromosome is duplicated to produce two closely paired sister DNA molecules (called sister chromatids). A cell with only one type of chromosome, present in maternal and paternal copies, 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 the members of each sister-chromatid 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.

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 4–19).

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. Eukaryotic chromosomes contain many origins of replication to ensure that the entire chromosome can be replicated rapidly, as discussed in detail in Chapter 5.

After DNA replication, the two sister chromatids that form each chromosome 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 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 eukaryotes, 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 a human centromere can contain up to a million nucleotide pairs and that it may not require a stretch of DNA with a defined nucleotide sequence. Instead, as we shall discuss later in this chapter, a human centromere is thought to consist of a large, regularly repeating protein–nucleic acid structure that can be inherited when a chromosome replicates.

Figure 4–18 A mitotic chromosome. A mitotic chromosome is a condensed duplicated chromosome in which the two new chromosomes, called sister chromatids, are still linked together (see Figure 4–17). The constricted region indicates the position of the centromere. (Courtesy of Terry D. Allen.)

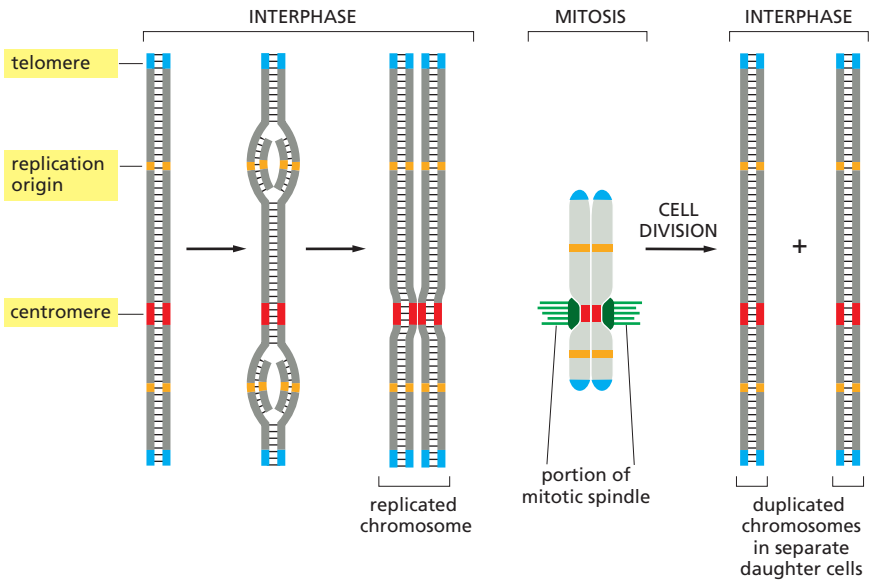


Figure 4–19 The three DNA sequences required to produce a eukaryotic chromosome that can be replicated and then segregated accurately at mitosis. Each chromosome has multiple origins of replication, one centromere, and two telomeres. Shown here is the sequence of events that a typical chromosome follows 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 a copy of the entire genome is distributed to each daughter cell during mitosis; the special structure that attaches the centromere to the spindle is a protein complex called the kinetochore *(dark green)*. 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.

DNA Molecules Are Highly Condensed in Chromosomes

All eukaryotic 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 over 7000-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.

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 extreme degree in the M phase of the cell cycle. Much less visible, but of enormous interest and importance, specific regions of interphase chromosomes decondense to allow 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 Eukaryotic Chromosome Structure

The proteins that bind to the DNA to form eukaryotic chromosomes are traditionally divided into two classes: the **histones** and the *non-histone chromosomal proteins*, each contributing about the same mass to a chromosome as the DNA. The complex of both classes of protein with the nuclear DNA of eukaryotic cells is known as **chromatin** (**Figure 4–20**).

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 appears to be in the form of a fiber with a diameter of about 30 nm (**Figure 4–21**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 4–21B). The string is DNA, and each bead is a "nucleosome core particle" that consists of DNA wound around a histone core (**Movie 4.2**).

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,

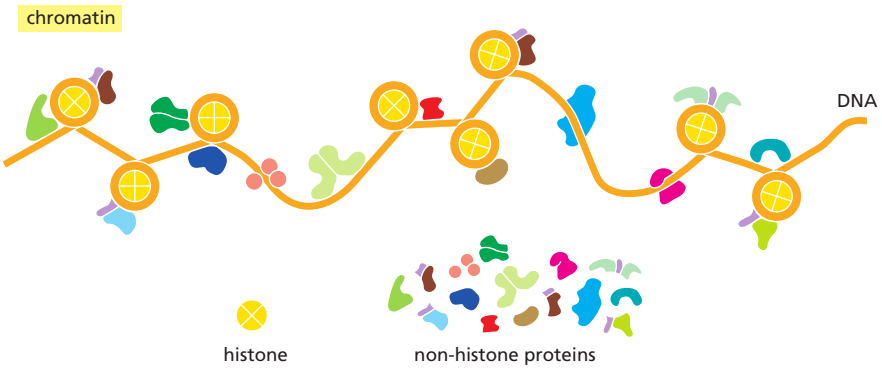


Figure 4–20 Chromatin. As illustrated, chromatin consists of DNA bound to both histone and non-histone proteins. The mass of histone protein present is about equal to the total mass of non-histone protein, but—as schematically indicated here—the latter class is composed of an enormous number of different species. In total, a chromosome is about one-third DNA and two-thirds protein by mass.

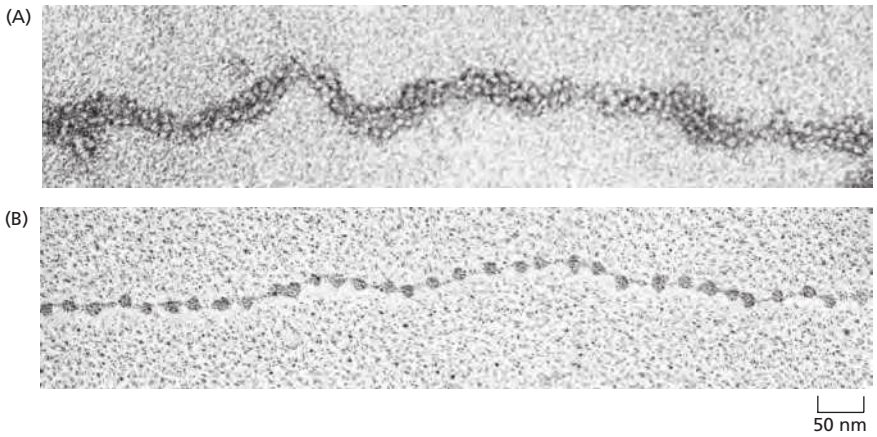


Figure 4–21 Nucleosomes as seen in the electron microscope. (A) Chromatin isolated directly from an interphase nucleus appears in the electron microscope as a thread about 30 nm thick. (B) This electron micrograph shows a length of chromatin that has been experimentally unpacked, or decondensed, after isolation to show the nucleosomes. (A, courtesy of Barbara Hamkalo; B, courtesy of Victoria Foe.)

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 4–22**).

The region of linker DNA that separates each nucleosome core particle from the next 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.

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 in a left-handed coil of 1.7 turns (**Figure 4–23**). 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 4–24**). 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.

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 sugar-phosphate backbone of the DNA. Numerous hydrophobic interactions and salt linkages also hold DNA and protein together in the nucleosome. 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 charges can effectively

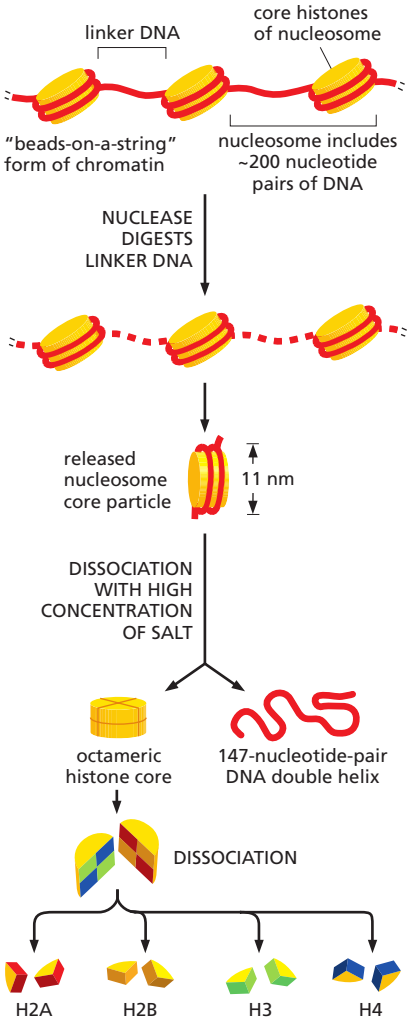


Figure 4–22 Structural organization of the nucleosome. A nucleosome contains a protein core made of eight histone molecules. In biochemical experiments, the nucleosome core particle can be released from isolated chromatin by digestion of the linker DNA with a nuclease, an enzyme that breaks down DNA. (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.

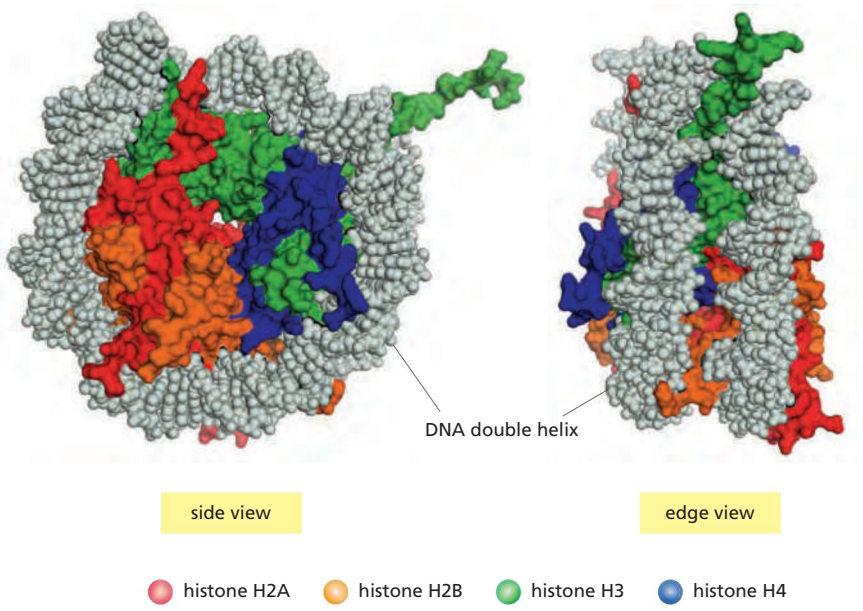


Figure 4–23 The structure of a nucleosome core particle, as determined by x-ray diffraction analyses of crystals. Each histone is colored according to the scheme in Figure 4–22, with the DNA double helix in *light gray*. (Adapted from K. Luger et al., *Nature* 389:251–260, 1997. With permission from Macmillan Publishers Ltd.)

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

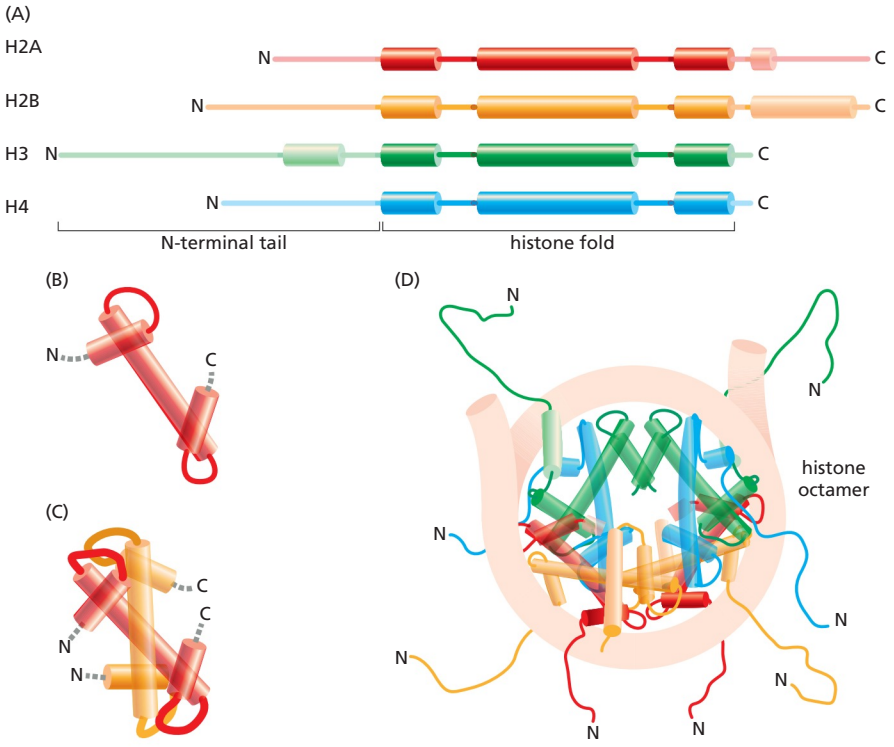


Figure 4–24 The overall structural organization of the core histones. (A) Each of the core histones contains an N-terminal tail, which is subject to several forms of covalent modification, and a histone fold region, as indicated. (B) The structure of the histone fold, which is formed by all four of the 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. (D) The final histone octamer on DNA. Note that all eight N-terminal tails of the histones protrude from the disc-shaped core structure. Their conformations are highly flexible, and they serve as binding sites for sets of other proteins.

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 4–25**). This probably explains some striking, but unusual, cases of very precise positioning of nucleosomes along a stretch of DNA. However, the sequence preference of nucleosomes must be weak 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 4–24D). 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 eukaryotic proteins. For example, the amino acid sequence of histone H4 from a pea differs from that of a cow 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. But in addition to this remarkable conservation, eukaryotic organisms also produce smaller amounts of specialized variant core histones that differ in amino acid sequence from the main ones. As discussed later, these variants, combined with the surprisingly large number of covalent modifications that can be added to the histones in nucleosomes, give rise to a variety of chromatin structures in cells.

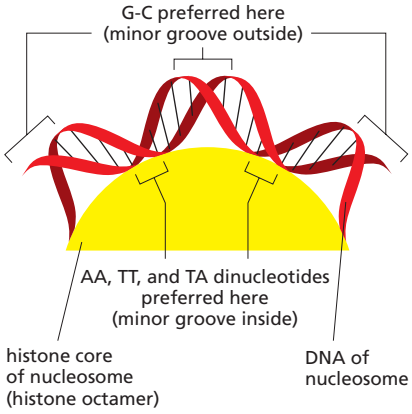


Figure 4–25 The bending of DNA in a nucleosome. The DNA helix makes 1.7 tight turns around the histone octamer. This diagram illustrates how the minor groove is compressed on the inside of the turn. Owing to structural features of the DNA molecule, the indicated dinucleotides are preferentially accommodated in such a narrow minor groove, which helps to explain why certain DNA sequences will bind more tightly than others to the nucleosome core.

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 would remain 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 easy access to many specific DNA sequences. It would also hinder 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 a rate of about four 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.

For the chromatin in a cell, a further loosening of DNA–histone contacts is clearly required, because eukaryotic cells contain a large variety of ATP-dependent *chromatin remodeling complexes*. These complexes include a subunit that hydrolyzes ATP (an ATPase evolutionarily related to the DNA helicases discussed in Chapter 5). This subunit 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, the protein complex changes the structure of a nucleosome temporarily, making the DNA less tightly bound to the histone core. Through repeated cycles of ATP hydrolysis that pull the nucleosome core along the DNA double helix, the remodeling complexes can catalyze *nucleosome sliding*. In this way, they can reposition nucleosomes to expose specific regions of DNA, thereby making them available to other proteins in the cell (**Figure 4–26**). In addition, by cooperating with a variety of other proteins that bind to histones and 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 4–27**). As a result of such processes, measurements reveal that a typical nucleosome is replaced on the DNA every one or two hours inside the cell.

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, some of which bind to specific modifications on histones (see Figure 4–26C). 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 4–40, below).

Although some DNA sequences bind more tightly than others to the nucleosome core (see Figure 4–25), 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 elsewhere. The exact positions of nucleosomes along a stretch of DNA therefore depend mainly on the presence and nature of other proteins bound to the DNA. And due to the presence of ATP-dependent chromatin 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 arrays in which the DNA is even more highly condensed. Thus, when nuclei are very gently lysed onto an electron microscope grid, much 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 4–21).

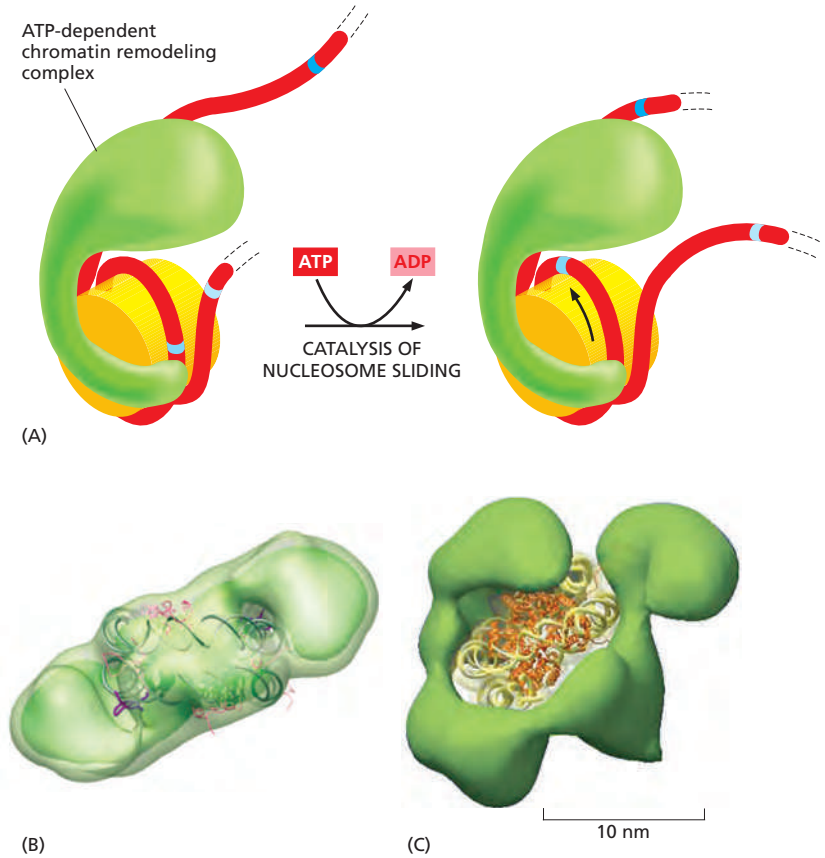


Figure 4–26 The nucleosome sliding catalyzed by ATP-dependent chromatin remodeling complexes. (A) Using the energy of ATP hydrolysis, the remodeling complex is thought to push on the DNA of its bound nucleosome and loosen its attachment to the nucleosome core. Each cycle of ATP binding, ATP hydrolysis, 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. (B) The structure of a nucleosome-bound dimer of the two identical ATPase subunits *(green)* that slide nucleosomes back and forth in the ISW1 family of chromatin remodeling complexes. (C) The structure of a large chromatin remodeling complex, showing how it is thought to wrap around a nucleosome. Modeled in *green* is the yeast RSC complex, which contains 15 subunits—including an ATPase and at least four subunits with domains that recognize specific covalently modified histones. (B, from L.R. Racki et al., *Nature* 462:1016–1021, 2009. With permission from Macmillan Publishers Ltd; C, adapted from A.E. Leschziner et al., *Proc. Natl Acad. Sci. USA* 104:4913–4918, 2007.)

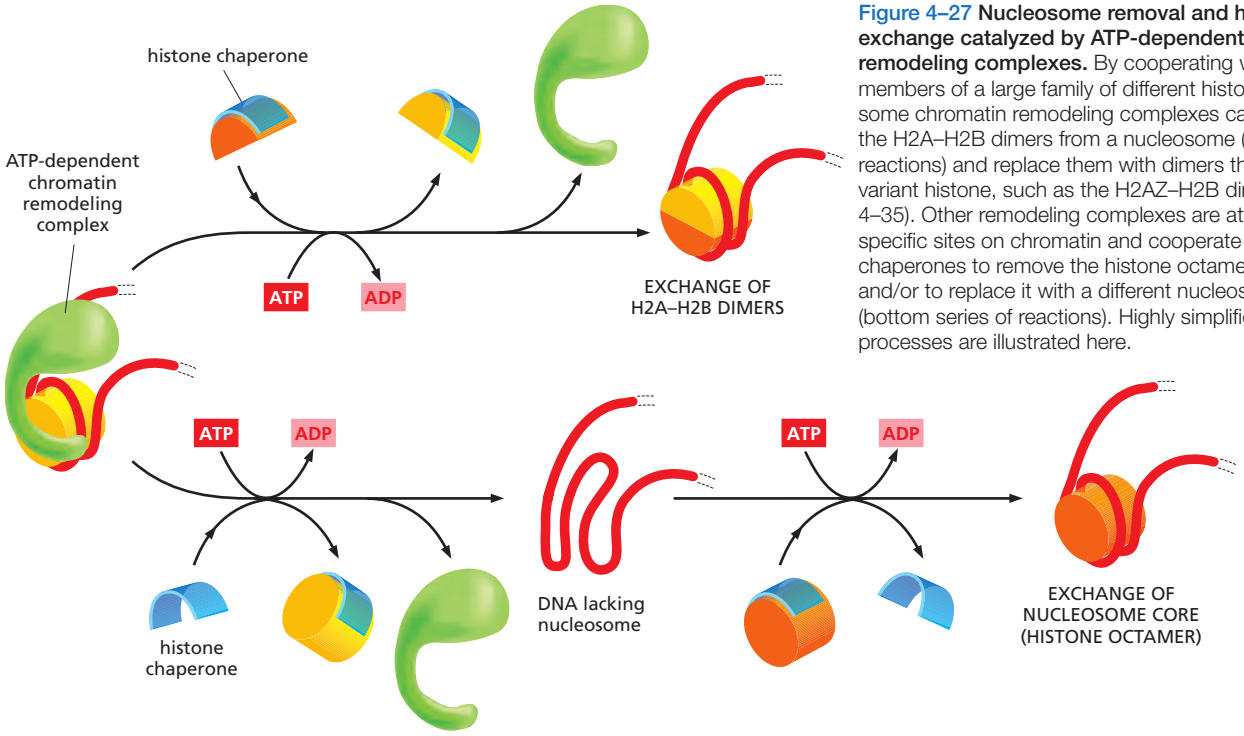


Figure 4–27 Nucleosome removal and histone exchange catalyzed by ATP-dependent chromatin remodeling complexes. By cooperating with specific members of a large family of different histone chaperones, some chromatin remodeling complexes can remove the H2A–H2B dimers from a nucleosome (top series of reactions) and replace them with dimers that contain a variant histone, such as the H2AZ–H2B dimer (see Figure 4–35). Other remodeling complexes are attracted to specific sites on chromatin and cooperate with histone chaperones to remove the histone octamer completely and/or to replace it with a different nucleosome core (bottom series of reactions). Highly simplified views of the processes are illustrated here.

How nucleosomes are organized into condensed arrays is unclear. The structure of a tetranucleosome (a complex of four nucleosomes) obtained by x-ray crystallography and high-resolution electron microscopy of reconstituted chromatin have been used to support a zigzag model for the stacking of nucleosomes in a 30-nm fiber (**Figure 4–28**). But cryoelectron microscopy of carefully prepared nuclei suggests that most regions of chromatin are less regularly structured.

What causes nucleosomes to stack so tightly on each other? Nucleosome-to-nucleosome linkages that involve histone tails, most notably the H4 tail, constitute one important factor (**Figure 4–29**). Another 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. This change in the exit path of DNA is thought to help compact nucleosomal DNA (**Figure 4–30**).

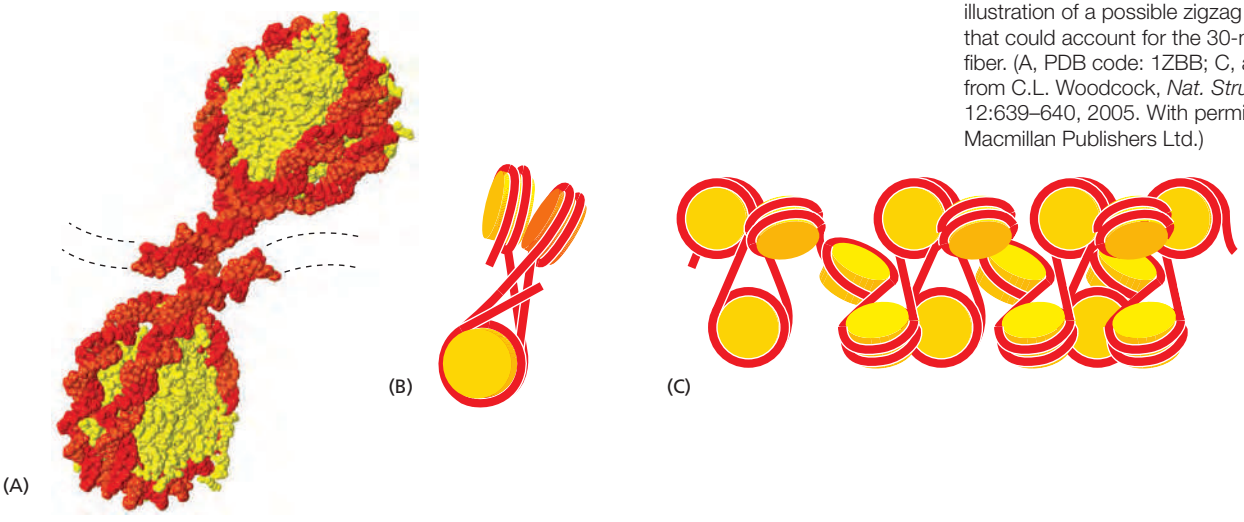


Figure 4–28 A zigzag model for the 30-nm chromatin fiber. (A) The conformation of two of the four nucleosomes in a tetranucleosome, from a structure determined by x-ray crystallography. (B) Schematic of the entire tetranucleosome; the fourth nucleosome is not visible, being stacked on the bottom nucleosome and behind it in this diagram. (C) Diagrammatic illustration of a possible zigzag structure that could account for the 30-nm chromatin fiber. (A, PDB code: 1ZBB; C, adapted from C.L. Woodcock, *Nat. Struct. Mol. Biol.* 12:639–640, 2005. With permission from Macmillan Publishers Ltd.)

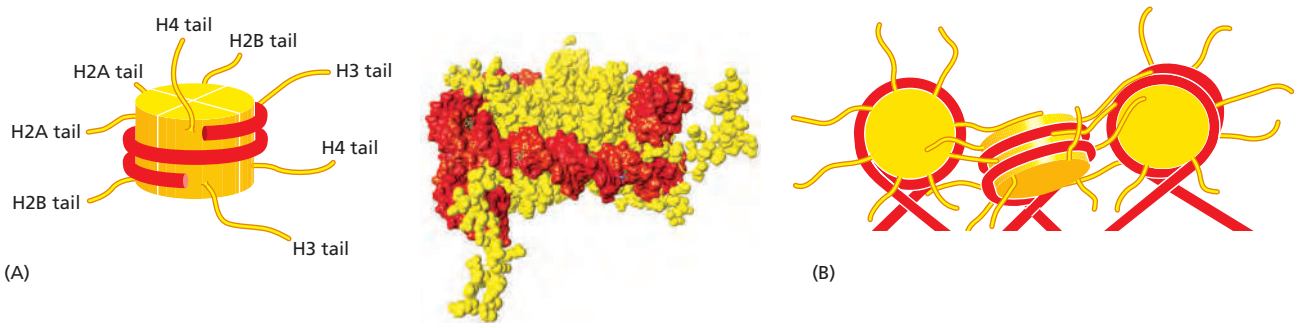


Figure 4–29 **A model for the role played by histone tails in the compaction of chromatin.** (A) A schematic diagram shows the approximate exit points of the eight histone tails, one from each histone protein, that extend from each nucleosome. The actual structure is shown to its right. In the high-resolution structure of the nucleosome, the tails are largely unstructured, suggesting that they are highly flexible. (B) As indicated, the histone tails are thought to be involved in interactions between nucleosomes that help to pack them together. (A, PDB code: 1KX5.)

Most eukaryotic organisms make several histone H1 proteins of related but quite distinct amino acid sequences. The presence of many other DNA-binding proteins, as well as proteins that bind directly to histones, is certain to 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 eukaryotes, 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, bound to a large set of proteins. The human genome contains 3.2 × 10^9 DNA nucleotide pairs, divided between 22 different autosomes (present in two copies each) 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 important nucleotide sequences: replication origins and telomeres allow the DNA molecule to be efficiently replicated, while a centromere attaches the sister DNA molecules to the mitotic spindle, ensuring their accurate segregation to daughter cells during the M phase of the cell cycle.

The DNA in eukaryotes 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. Even though compact, the structure of chromatin 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.

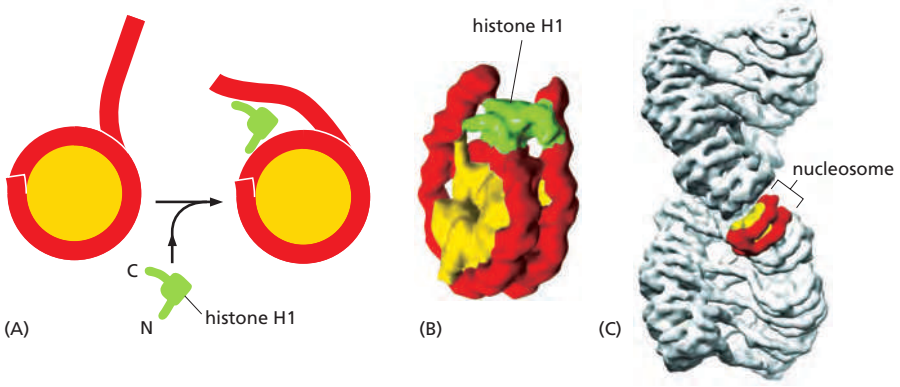


Figure 4–30 **How the linker histone binds to the nucleosome.** The position and structure of histone H1 is shown. The H1 core region constrains an additional 20 nucleotide pairs of DNA where it exits from the nucleosome core and is important for compacting chromatin. (A) Schematic, and (B) structure inferred for a single nucleosome from a structure determined by high-resolution electron microscopy of a reconstituted chromatin fiber (C). (B and C, adapted from F. Song et al., *Science* 344:376–380, 2014.)

CHROMATIN STRUCTURE AND FUNCTION

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. Mechanisms of this type have a variety of important functions in cells. Most strikingly, certain types of chromatin structure can be inherited; that is, the structure can be directly passed down from a cell to its descendants. Because the cell memory that results is based on an inherited chromatin 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.

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 different mechanisms that can produce it. Here, we are concerned with only one, that based on chromatin structure. We begin this section by reviewing the observations that first demonstrated that chromatin structures can be inherited. We then describe some of the chemistry that makes this possible—the covalent modification of histones in nucleosomes. These modifications have many functions, inasmuch as they serve as recognition sites for protein domains that link specific protein complexes to different regions of chromatin. Histones thereby have effects on gene expression, as well as on many other DNA-linked processes. Through such mechanisms, chromatin structure plays an important role in the development, growth, and maintenance of all eukaryotic organisms, including ourselves.

Heterochromatin Is Highly Organized and Restricts Gene Expression

Light-microscope studies in the 1930s distinguished two types of chromatin in the interphase nuclei of many higher eukaryotic 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 its molecular properties. It is highly concentrated in certain specialized regions, most notably at the centromeres and telomeres introduced previously (see Figure 4–19), but it is also present at many other locations along chromosomes—locations that can vary according to the physiological state of the cell. In a typical mammalian cell, more than 10% of the genome is packaged in this way.

The DNA in heterochromatin typically contains few genes, and when euchromatic regions are converted to a heterochromatic state, their genes are generally switched off as a result. However, we know now that the term *heterochromatin* encompasses several distinct modes of chromatin compaction that have different implications for gene expression. Thus, heterochromatin should not be thought of as simply encapsulating "dead" DNA, but rather as a descriptor for compact chromatin domains that share the common feature of being unusually resistant to gene expression.

The Heterochromatic State Is Self-Propagating

Through chromosome breakage and rejoining, whether brought about by a natural genetic accident or by experimental artifice, a piece of chromosome that is normally euchromatic can be translocated into the neighborhood of heterochromatin. Remarkably, this often causes *silencing*—inactivation—of the normally active genes. This phenomenon is referred to as a *position effect*. It reflects a spreading of the heterochromatic state into the originally euchromatic region, and it has provided important clues to the mechanisms that create and maintain heterochromatin. First recognized in *Drosophila*, position effects have now been observed in many eukaryotes, including yeasts, plants, and humans.

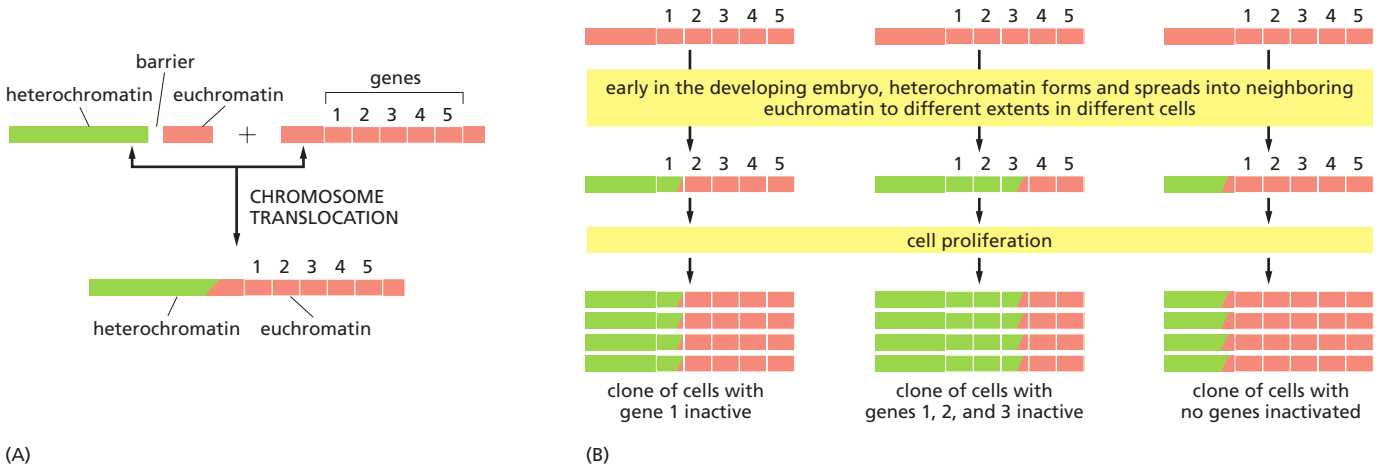


Figure 4–31 The cause of position effect variegation in *Drosophila*. (A) Heterochromatin *(green)* is normally prevented from spreading into adjacent regions of euchromatin *(red)* by *barrier* DNA sequences, which we shall discuss shortly. In flies that inherit certain chromosomal rearrangements, however, 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 subsequently 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 4–32). 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, the condensation of DNA into heterochromatin can "skip over" some regions of chromatin, sparing the genes that lie within them from repressive effects.

In chromosome breakage-and-rejoining events of the sort just described, the zone of silencing, where euchromatin is converted to a heterochromatic state, spreads for different distances in different early cells in the fly embryo. Remarkably, these differences then are perpetuated for the rest of the animal's life: in each cell, once the heterochromatic condition is established on a piece of chromatin, it tends to be stably inherited by all of that cell's progeny (**Figure 4–31**). This remarkable phenomenon, called **position effect variegation**, was first recognized through a detailed genetic analysis of the mottled loss of red pigment in the fly eye (**Figure 4–32**). It shares features with the extensive spread of heterochromatin that inactivates one of the two X chromosomes in female mammals. There too, a random process acts in each cell of the early embryo to dictate which X chromosome will be inactivated, and that same X chromosome then remains inactive in all the cell's progeny, creating a mosaic of different clones of cells in the adult body (see Figure 7–50).

These observations, taken together, point to a fundamental strategy of heterochromatin formation: heterochromatin begets more heterochromatin. This positive feedback can operate both in space, causing the heterochromatic state to spread along the chromosome, and in time, across cell generations, propagating the heterochromatic state of the parent cell to its daughters. The challenge is to explain the molecular mechanisms that underlie this remarkable behavior.

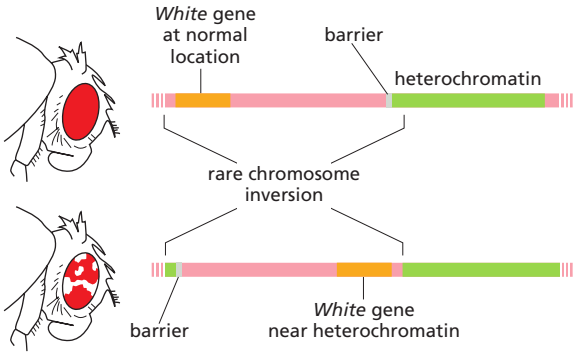
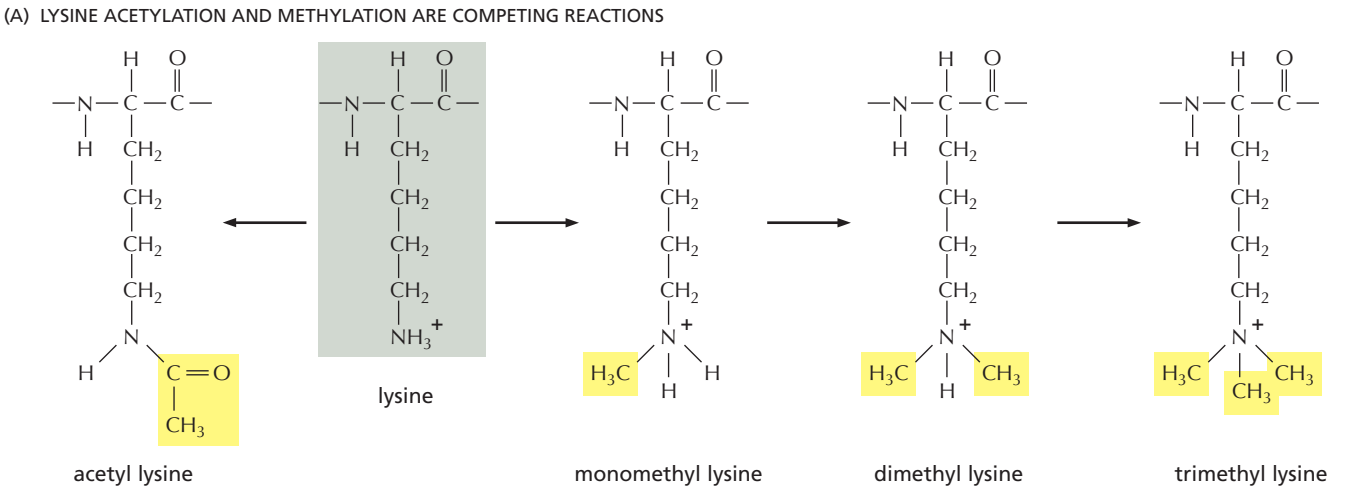


Figure 4–32 The discovery of position effects on gene expression. The *White* gene in the fruit fly *Drosophila* controls eye pigment production and is named after the mutation that first identified it. Wild-type flies with a normal *White* gene (*White*$^+$) have normal pigment production, which gives 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 4–31). 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.



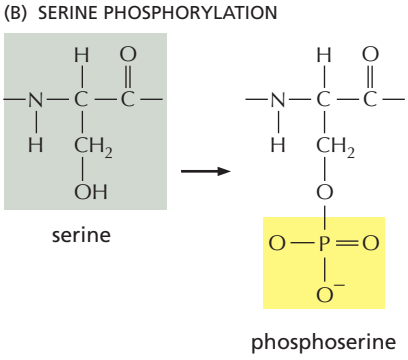


Figure 4–33 Some prominent types of covalent amino acid side-chain modifications found on nucleosomal histones. (A) Three different levels of lysine methylation are shown; each can be recognized by a different binding protein and thus each can have a different significance for the cell. Note that acetylation 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 of histones not shown here include the mono- or dimethylation 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.

As a first step, one can carry out a search for the molecules that are involved. This has been done by means of *genetic screens*, in which large numbers of mutants are generated, after which one picks out those that show an abnormality of the process in question. Extensive genetic screens in *Drosophila*, fungi, and mice have identified more than 100 genes whose products either enhance or suppress the spread of heterochromatin and its stable inheritance—in other words, genes that serve as either enhancers or suppressors of position effect variegation. Many of these genes turn out to code for non-histone chromosomal proteins that interact with histones and are involved in modifying or maintaining chromatin structure. We shall discuss how they work in the sections that follow.

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 trimethylation of lysines, and the phosphorylation of serines (**Figure 4–33**). A large number of these side-chain modifications occur on the eight relatively unstructured N-terminal "histone tails" that protrude from the nucleosome (**Figure 4–34**). However, there are also more than 20 specific side-chain modifications on the nucleosome's globular core.

All of the above types of modifications are reversible, with one enzyme serving to create a particular type of modification, and another to remove it. These enzymes are highly specific. 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 depends on *transcription regulator proteins* (sometimes called "transcription factors"). As we shall explain in Chapter 7, these proteins recognize and bind to specific DNA sequences in the chromosomes. They are produced at

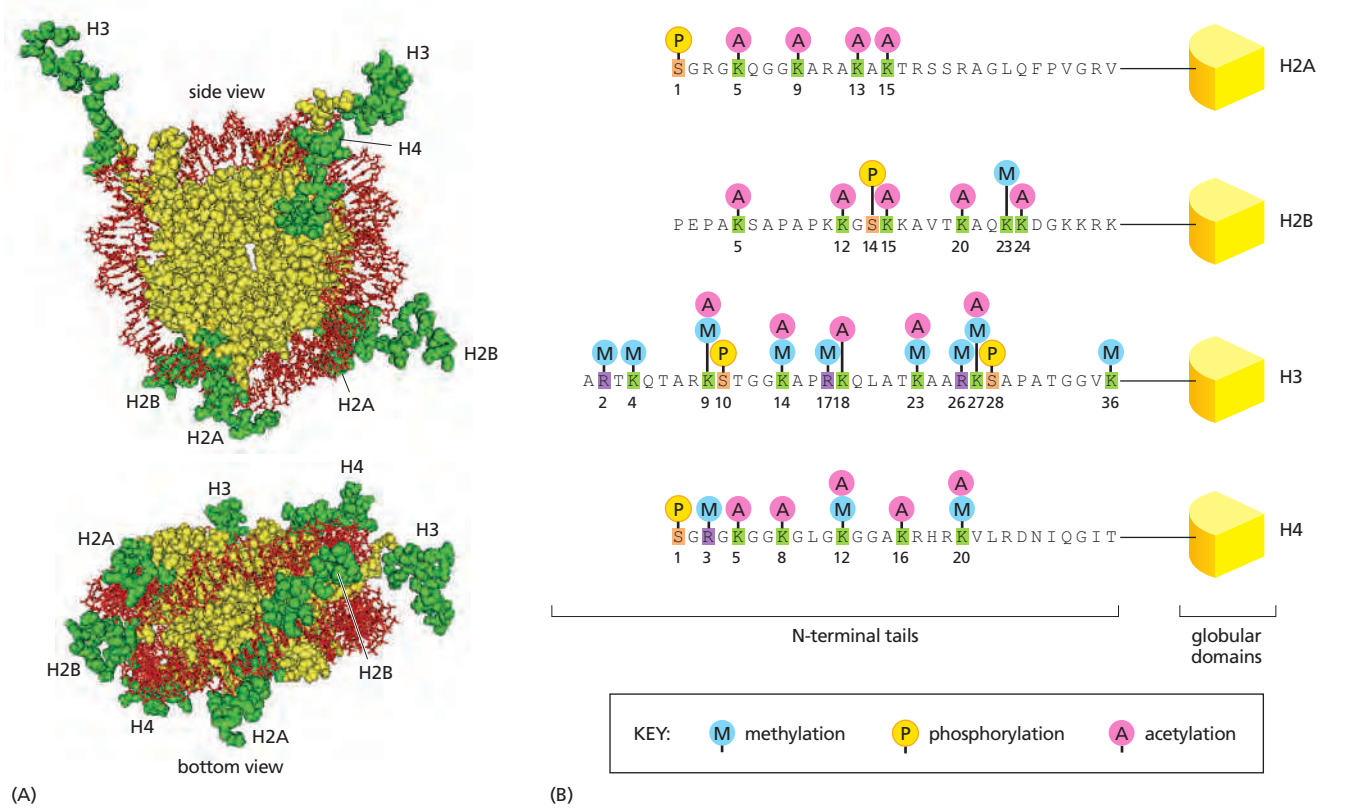


Figure 4–34 The covalent modification of core histone tails. (A) The structure of the nucleosome highlighting the location of the first 30 amino acids in each of its eight N-terminal histone tails *(green)*. These tails are unstructured and highly mobile, and thus will change their conformation depending on other bound proteins. (B) Well-documented modifications of the four histone core proteins are indicated. Although only a single symbol is used here for methylation (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 3–69). Not shown are more than 20 possible modifications located in the globular core of the histones. (A, PDB: 1KX5; B, adapted from H. Santos-Rosa and C. Caldas, *Eur. J. Cancer* 41:2381–2402, 2005. With permission from Elsevier.)

different times and places in the life of an organism, thereby determining where and when the chromatin-modifying enzymes will act. In this way, the DNA sequence ultimately determines how histones are modified. But in at least some cases, the covalent modifications on nucleosomes can persist long after the transcription regulator proteins that first induced them have disappeared, thereby providing the cell with a memory of its developmental history. Most remarkably, as in the related phenomenon of position effect variegation discussed above, this memory can be transmitted from one cell generation to the next.

Very different patterns of covalent modification are found on different groups of nucleosomes, depending both on their exact position in the genome and on the history of the cell. The modifications of the histones are carefully controlled, and they have important consequences. The acetylation of lysines on the N-terminal tails loosens chromatin structure, in part because adding an acetyl group to lysine removes its positive charge, thereby reducing the affinity of the tails for adjacent nucleosomes. However, the most profound effects of the histone modifications lie in their ability to recruit specific other proteins to the modified stretch of chromatin. Trimethylation of one specific lysine on the histone H3 tail, for instance, attracts the heterochromatin-specific protein HP1 and contributes to the establishment and spread of heterochromatin. More generally, the recruited proteins act with the modified histones to determine how and when genes will be expressed, as well as other chromosome functions. In this way, the precise structure of each domain of chromatin governs the readout of the genetic information that it contains, and thereby the structure and function of the eukaryotic cell.

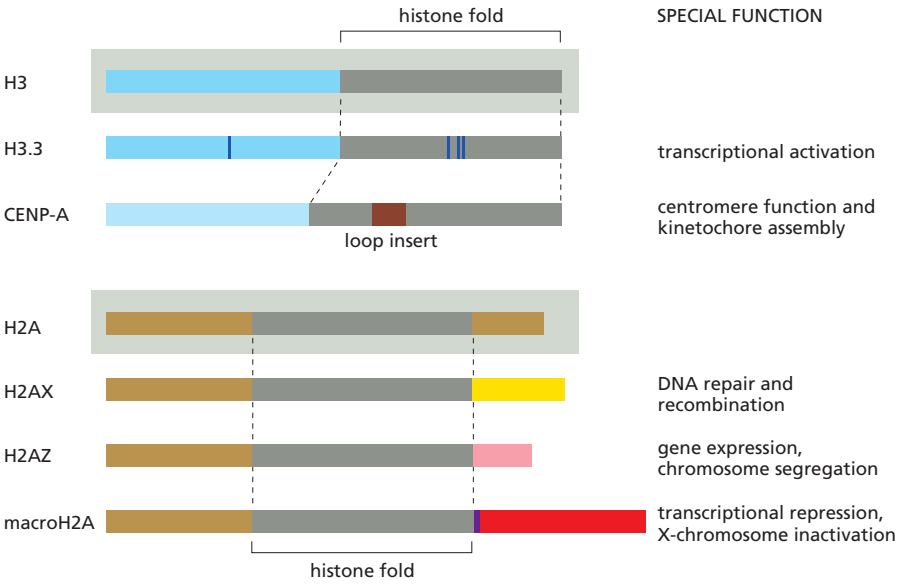


Figure 4–35 The structure of some histone variants compared with the major histone that they replace. The histone variants are inserted into nucleosomes at specific sites on chromosomes by ATP-dependent chromatin remodeling enzymes that act in concert with histone chaperones (see Figure 4–27). The CENP-A (Centromere Protein-A) variant of histone H3 is discussed later in this chapter (see Figure 4–42); other variants are discussed in Chapter 7. The sequences in each variant that are colored differently (compared to the major histone above it) denote regions with an amino acid sequence different from this major histone. (Adapted from K. Sarma and D. Reinberg, *Nat. Rev. Mol. Cell Biol.* 6:139–149, 2005. With permission from Macmillan Publishers Ltd.)

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

In addition to the four highly conserved standard core histones, eukaryotes also contain a few variant histones that can also 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. Variants are known for each of the core histones with the exception of H4; some examples are shown in **Figure 4–35**.

The major histones are synthesized primarily during the S phase of the cell cycle and assembled into nucleosomes on the daughter DNA helices just behind the replication fork (see Figure 5–32). 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 4–27).

Covalent Modifications and Histone Variants Act in Concert to Control Chromosome Functions

The number of possible distinct markings on an individual nucleosome is in principle enormous, and this potential for diversity is still greater when we allow for nucleosomes that contain histone variants. However, the histone modifications are known to occur in coordinated sets. More than 15 such sets can be identified in mammalian cells. However, it is not yet clear how many different types of chromatin are functionally important in cells.

Some combinations are known to have a specific meaning for the cell in the sense that they determine how and when the DNA packaged in the nucleosomes is to be accessed or manipulated—a fact that led to the idea of a "*histone code.*" 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 others signal when and how gene expression should take place. Various regulatory proteins contain small domains that bind to specific marks, recognizing, for example, a trimethylated lysine 4 on histone H3 (**Figure 4–36**). These domains are often linked together as modules in a single large

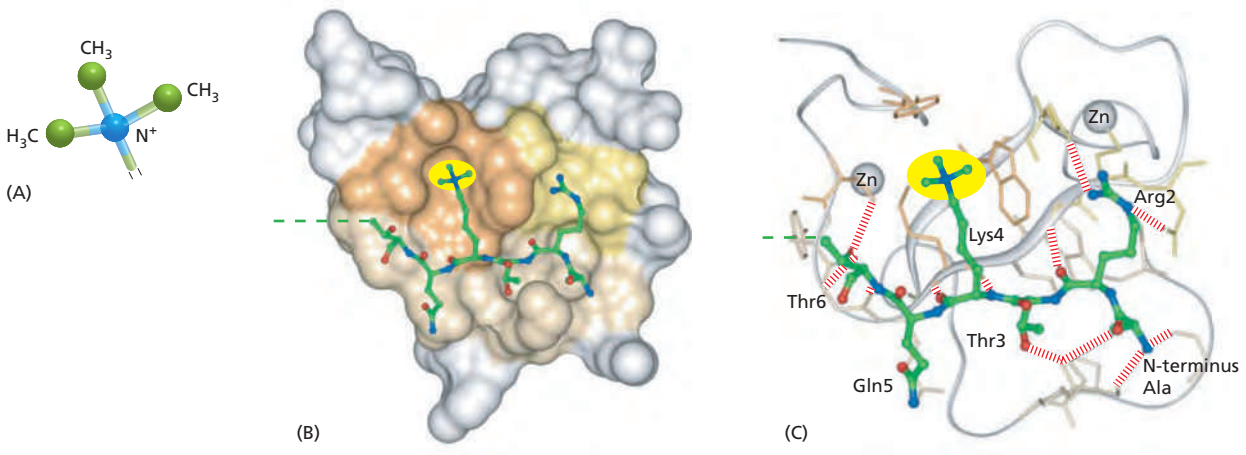


Figure 4–36 How a mark on a nucleosome is read. The figure shows the structure of a protein module (called an ING PHD domain) that specifically recognizes histone H3 trimethylated on lysine 4. (A) A trimethyl group. (B) Space-filling model of an ING PHD domain bound to a histone tail (*green*, with the trimethyl group highlighted in *yellow*). (C) A ribbon model showing how the N-terminal six amino acids in the H3 tail are recognized. The *red lines* represent hydrogen bonds. This is one of a family of PHD domains that recognize methylated lysines on histones; different members of the family bind tightly to lysines located at different positions, and they can discriminate between a mono-, di-, and trimethylated 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. Peña et al., *Nature* 442:100–103, 2006. With permission from Macmillan Publishers Ltd.)

protein or protein complex, which thereby recognizes a specific combination of histone modifications (**Figure 4–37**). The result is a *reader complex* that allows particular combinations of markings on chromatin to attract additional proteins, so as to execute an appropriate biological function at the right time (**Figure 4–38**).

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 that can be readily altered as a cell's needs change. Although much remains to be learned about the meaning of the different histone modifications, a few well-studied examples of the information that can be encoded in the histone H3 tail are listed in **Figure 4–39**.

A Complex of Reader and Writer Proteins Can Spread Specific Chromatin Modifications Along a Chromosome

The phenomenon of position effect variegation described previously requires that some modified forms of chromatin have the ability to spread for substantial distances along a chromosomal DNA molecule (see Figure 4–31). How is this possible?

The enzymes that add or remove modifications to 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 (transcription regulators) discussed in Chapters 6 and 7 (for a specific example,

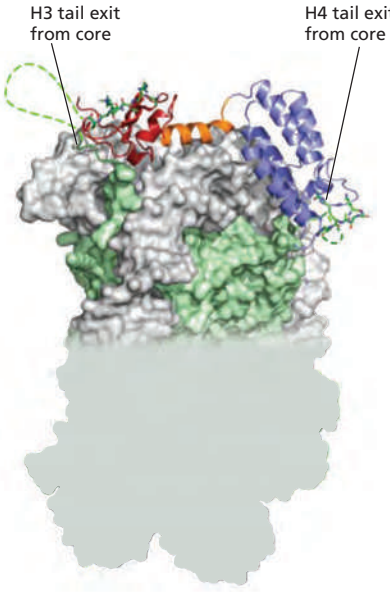


Figure 4–37 Recognition of a specific combination of marks on a nucleosome. In the example shown, two adjacent domains that are part of the NURF (Nucleosome Remodeling Factor) chromatin remodeling complex bind to the nucleosome, with the PHD domain *(red)* recognizing a methylated H3 lysine 4 and another domain (a bromodomain, *blue*) recognizing an acetylated H4 lysine 16. These two histone marks constitute a unique histone modification pattern that occurs in subsets of nucleosomes in human cells. Here the two histone tails are indicated by *green dotted* lines, and only half of one nucleosome is shown. (Adapted from A.J. Ruthenburg et al., *Cell* 145:692–706, 2011. With permission from Elsevier.)

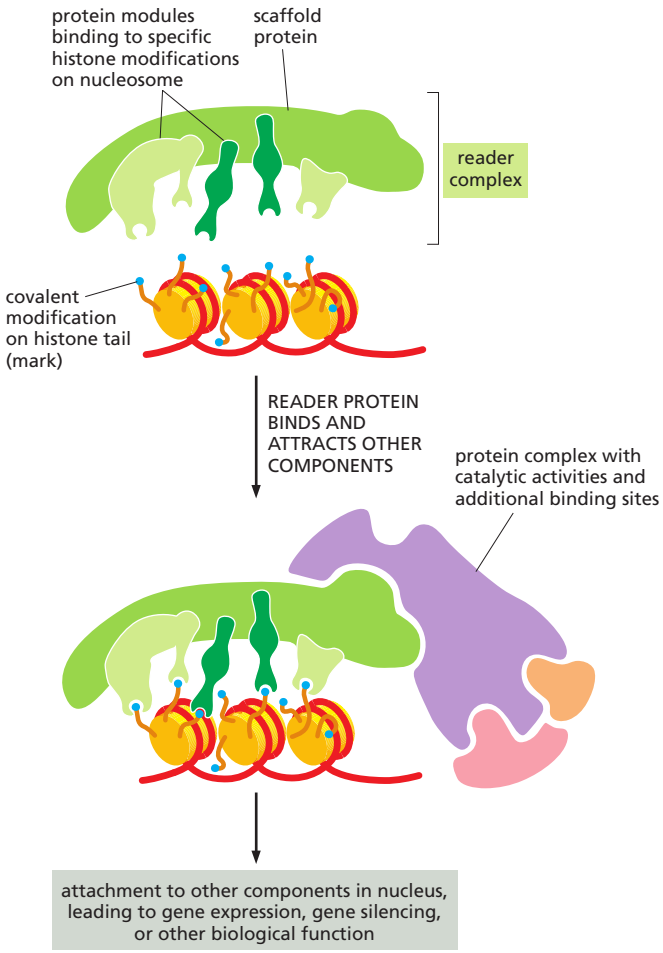


Figure 4–38 Schematic diagram showing how a particular combination of histone modifications can be recognized by a reader complex. A large protein complex that contains a series of protein modules, each of which recognizes a specific histone mark, is schematically illustrated *(green)*. This "reader 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 the additional protein complexes *(purple)* needed to catalyze a biological function.

see Figure 7–20). But after a modifying enzyme "writes" its mark on one or a few neighboring nucleosomes, events that resemble a chain reaction can ensue. In such a case, the "writer enzyme" works in concert with a "reader protein" located in the same protein complex. The reader protein contains a module that recognizes the mark and binds tightly to the newly modified nucleosome (see Figure

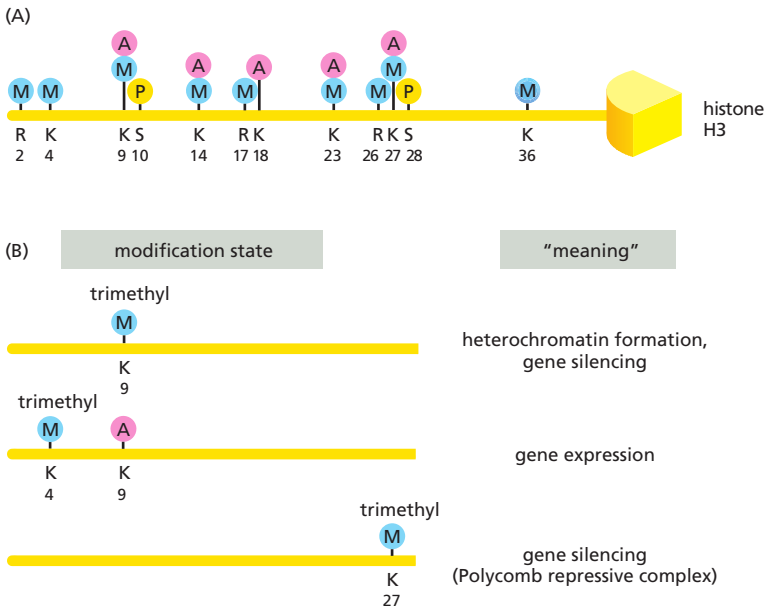


Figure 4–39 Some specific meanings of histone modifications. (A) The modifications on the histone H3 N-terminal tail are shown, repeated from Figure 4–34. (B) The H3 tail can be marked by different sets of modifications that act in combination to convey a specific meaning. Only a small number of the meanings are known, including the three examples shown. Not illustrated is the fact that, as just implied (see Figure 4–38), reading a histone mark 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 trimethyl groups) are generally required. Thus, for 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 4–40). Also important in this process, however, is a synergistic trimethylation of the histone H4 N-terminal tail on lysine 20.

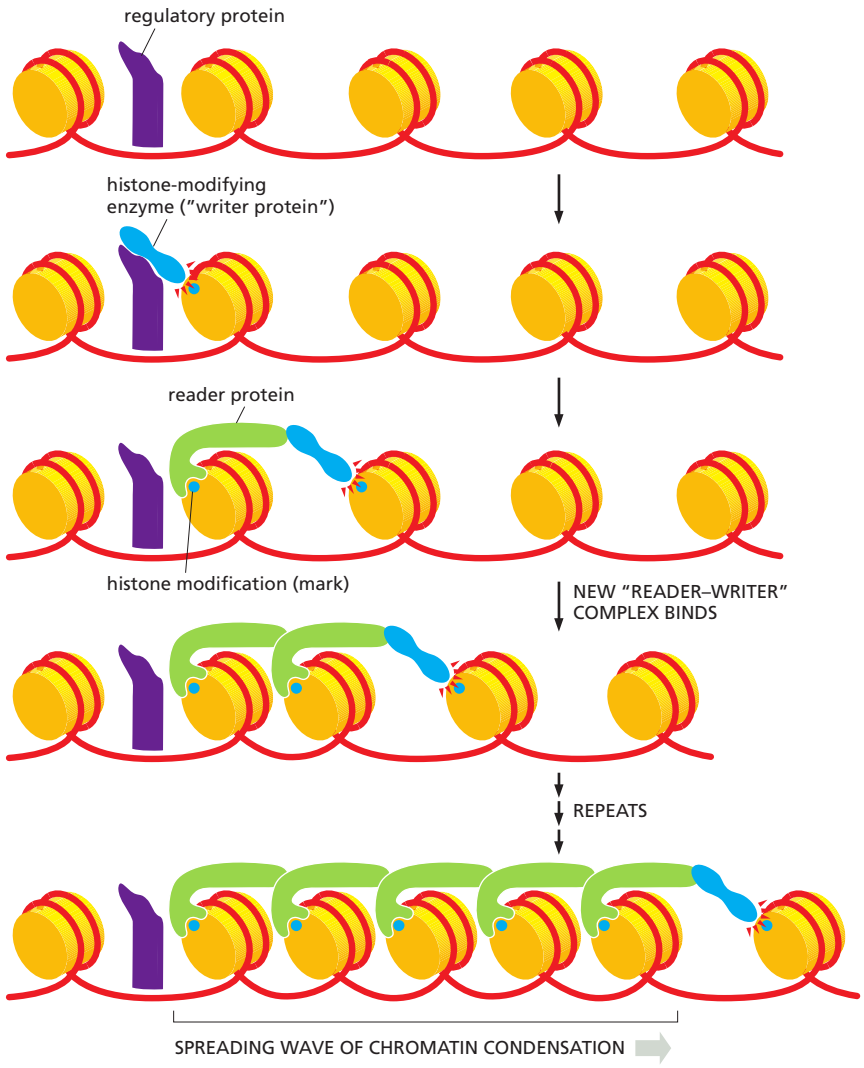


Figure 4–40 How the recruitment of a reader–writer complex can spread chromatin changes along a chromosome. The writer is an enzyme that creates a specific modification on one or more of the four nucleosomal histones. After its recruitment to a specific site on a chromosome by a transcription regulatory protein, the writer collaborates with a 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; its binding to that mark can be shown to activate the writer. In this schematic example, a spreading wave of chromatin condensation is thereby induced. Not shown are the additional proteins involved, including an ATP-dependent chromatin remodeling complex required to reposition the modified nucleosomes.

4–36), activating an attached writer enzyme and positioning it 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 4–40**).

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 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 (see Figure 4–26C), and the reader, writer, and remodeling proteins can work in concert to either decondense or condense long stretches of chromatin as the reader moves progressively along the nucleosome-packaged DNA.

A similar process is used to remove histone modifications from specific regions of the DNA; in this case, an "eraser enzyme," such as a histone demethylase or histone deacetylase, is recruited to the complex. As for the writer complex in Figure 4–40, sequence-specific DNA-binding proteins (transcription regulators) direct where such modifications occur (discussed in Chapter 7).

Some idea of the complexity of the above processes can be derived from the results of genetic screens for genes that either enhance or suppress the spreading and stability of heterochromatin, as manifest in effects on position effect variegation in *Drosophila* (see Figure 4–32). As pointed out previously, more than 100 such genes are known, and most of them are likely to code for 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 contains 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: certain DNA sequences mark the boundaries of chromatin domains and separate one such domain from another (see Figure 4–31). Several such *barrier sequences* have now been identified and characterized through the use of genetic engineering techniques that allow specific DNA segments to be deleted from, or inserted in, chromosomes.

For example, in cells that are destined to give rise to red blood cells, a sequence called HS4 normally separates the active chromatin domain that contains the human β-globin locus from an adjacent region of silenced, condensed chromatin. 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 position effect variegation similar to that observed in *Drosophila*. As described in Chapter 7, the consequences are dire: the globin genes are poorly expressed, and individuals who carry such a deletion have a severe form of anemia.

In genetic engineering experiments, the HS4 sequence is often added to both ends of a gene that is to be 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, and specific lysine methylations are required to spread heterochromatin, histone acetylases are logical candidates for the formation of DNA barriers to spreading (**Figure 4–41**). However, several other types of chromatin modifications are known that can also protect genes from silencing.

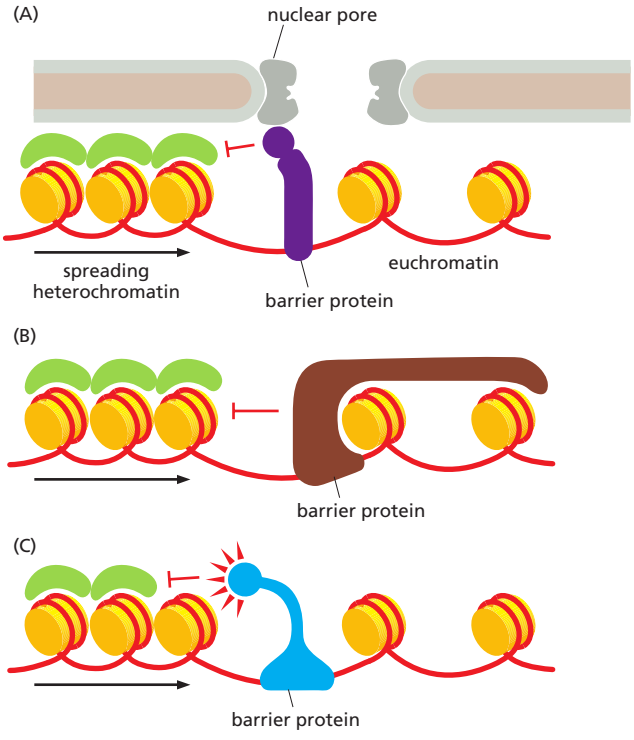


Figure 4–41 Some mechanisms of barrier action. These models are derived from experimental analyses of barrier action, and a combination of several of them may function at any one site. (A) The tethering of a region of chromatin to a large fixed site, such as the nuclear pore complex illustrated here, can form a barrier that stops the spread of heterochromatin. (B) The tight binding of barrier proteins to a group of nucleosomes can make this chromatin resistant to 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 binding of the HP1 protein needed to form a major form of heterochromatin. (Based on A.G. West and P. Fraser, *Hum. Mol. Genet.* 14:R101–R111, 2005. With permission from Oxford University Press.)

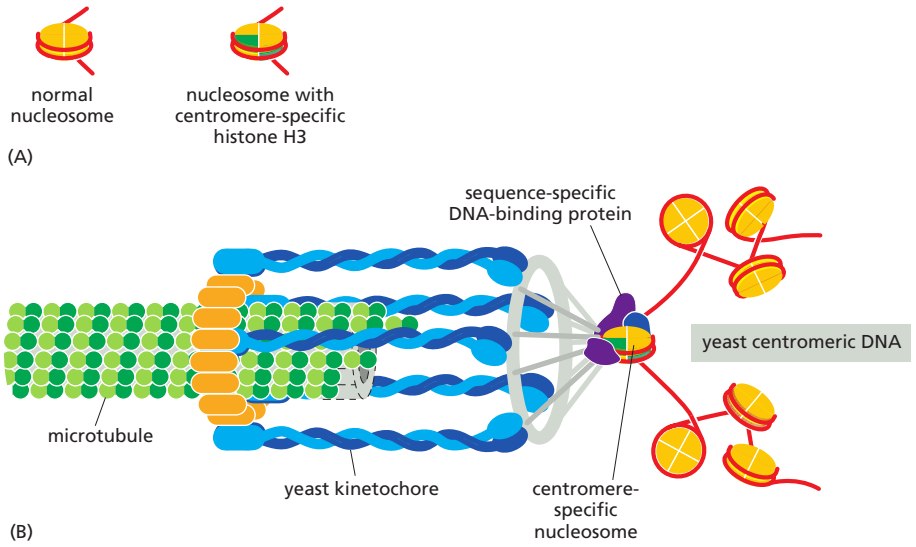


Figure 4–42 A model for the structure of a simple centromere. (A) In the yeast *Saccharomyces cerevisiae*, a special centromeric DNA sequence assembles a single nucleosome in which two copies of an H3 variant histone (called CENP-A in most organisms) replace the normal H3. (B) How peptide sequences unique to this variant histone (see Figure 4–35) help to assemble additional proteins, some of which form a kinetochore. The yeast 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 4–43). The kinetochore is discussed in detail in Chapter 17. (Adapted from A. Joglekar et al., *Nat. Cell Biol.* 8:581–585, 2006. With permission from Macmillan Publishers Ltd.)

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

Nucleosomes carrying histone variants have a distinctive character and are thought to be able to produce marks in chromatin that are unusually long-lasting. An important example is seen in the formation and inheritance of the specialized chromatin structure at the centromere, the region of each chromosome required for attachment to the mitotic spindle and orderly segregation of the duplicated copies of the genome into daughter cells each time a cell divides. In many complex organisms, including humans, each centromere is embedded in a stretch of special *centromeric chromatin* that persists throughout interphase, even though the centromere-mediated attachment to the spindle and movement of DNA occur only during mitosis. This chromatin contains a centromere-specific variant H3 histone, known as CENP-A (Centromere Protein-A; see Figure 4–35), plus additional proteins that pack the nucleosomes into particularly dense arrangements and form the kinetochore, the special structure required for attachment of the mitotic spindle (see Figure 4–19).

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 4–42**).

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, while they contain CENP-A, they 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 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 4–43**). It seems that centromeres in complex organisms are defined by an assembly of proteins, rather than by a specific DNA sequence.

Inactivation of some centromeres and genesis of others *de novo* seem to have played an essential part in evolution. Different species, even when quite closely

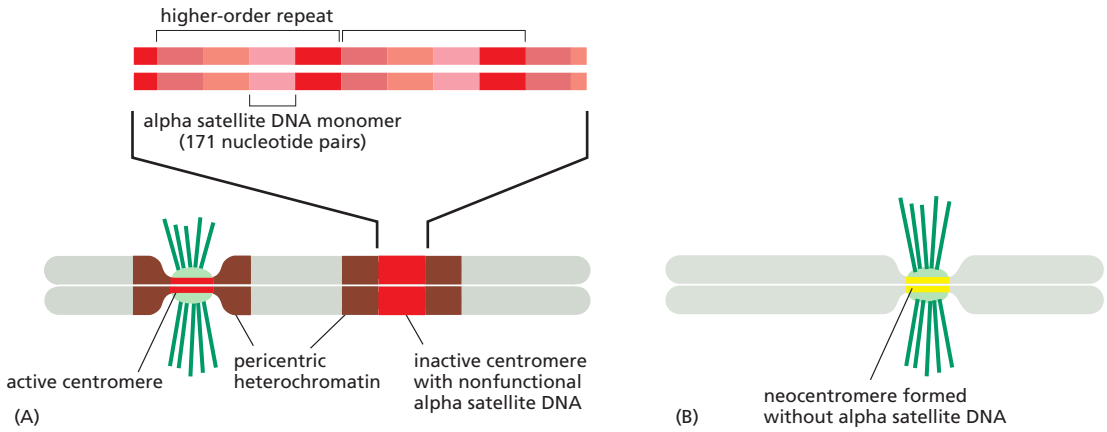


Figure 4–43 Evidence for the plasticity of human centromere formation. (A) A series of A-T-rich alpha satellite DNA sequences is repeated many thousands of times at each human centromere *(red)*, and is surrounded by *pericentric heterochromatin (brown).* However, due to an ancient chromosome breakage-and-rejoining event, some human chromosomes contain two blocks of alpha satellite DNA, each of which presumably functioned as a centromere in its original chromosome. Usually, chromosomes with two functional centromeres 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 become 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.

The complexity of centromeric chromatin is not illustrated in these diagrams. The alpha satellite DNA that forms centromeric chromatin in humans is packaged into alternating blocks of chromatin. One block is formed from a long string of nucleosomes containing the CENP-A H3 variant histone; the other block contains nucleosomes that are specially marked with dimethyl lysine 4 on the normal H3 histone. Each block is more than a thousand nucleosomes long. This centromeric chromatin is flanked by pericentric heterochromatin, as shown. The pericentric chromatin contains methylated lysine 9 on its H3 histones, along with HP1 protein, and it is an example of "classical" heterochromatin (see Figure 4–39).

related, often have different numbers of chromosomes; see Figure 4–14 for an extreme example. As we shall discuss below, detailed genome comparisons show that in many cases the changes in chromosome numbers have arisen through chromosome breakage-and-rejoining events, creating novel chromosomes, some of which must initially have contained abnormal numbers of centromeres—either more than one, or none at all. Yet stable inheritance requires that each chromosome should contain one centromere, and one only. It seems that surplus centromeres must have been inactivated, and/or new centromeres created, so as to allow the rearranged chromosome sets to be stably maintained.

Some Chromatin Structures Can Be Directly Inherited

The changes in centromere activity just discussed, once established, need to be perpetuated through subsequent cell generations. What could be the mechanism of this type of epigenetic inheritance?

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 sister DNA helices at a replication fork (see Figure 5–32). 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. One need only assume that the presence of the CENP-A histone in an inherited nucleosome selectively recruits more CENP-A histone to its newly formed neighbors.

There are some striking similarities between the formation and maintenance of centromeres and the formation and maintenance of some other regions of

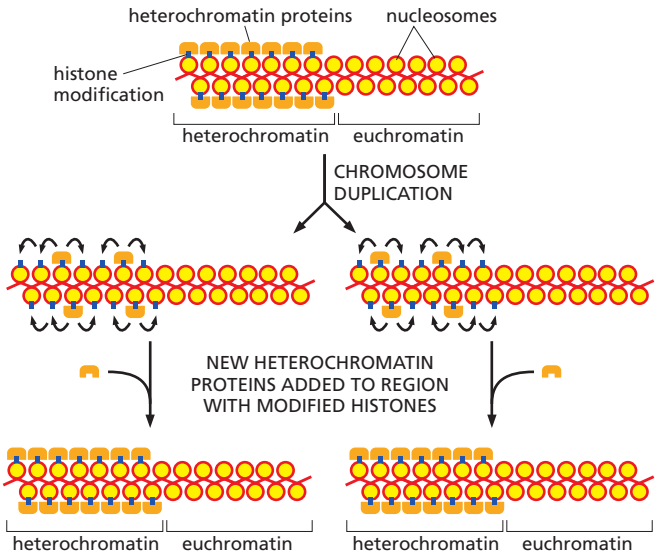


Figure 4–44 How the packaging of DNA in chromatin can be inherited following chromosome replication. In this model, some of the specialized chromatin components are distributed to each sister chromosome after DNA duplication, along with the specially marked nucleosomes that they bind. After DNA replication, 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 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 reader–writer–remodeling complexes operating in a manner similar to that previously illustrated in Figure 4–40.

heterochromatin. In particular, the entire centromere forms as an all-or-none entity, suggesting that the creation of centromeric chromatin is a highly cooperative process, spreading out from an initial seed in a manner reminiscent of the phenomenon of position effect variegation that we discussed earlier. In both cases, a particular chromatin structure, once formed, seems to be directly inherited on the DNA following each round of chromosome replication. A cooperative recruitment of proteins, along with the action of reader–writer complexes, can thus not only account for the spreading of specific forms of chromatin in space along the chromosome, but also for its propagation across cell generations—from parent cell to daughter cell (**Figure 4–44**).

Experiments with Frog Embryos Suggest that both Activating and Repressive Chromatin Structures Can Be Inherited Epigenetically

Epigenetic inheritance plays a central part in the creation of multicellular organisms. Their differentiated cell types become established during development, and persist thereafter even through repeated cell-division cycles. The daughters of a liver cell persist as liver cells, those of an epidermal cell as epidermal cells, and so on, even though they all contain the same genome; and this is because distinctive patterns of gene expression are passed on faithfully from parent cell to daughter cell. Chromatin structure has a role in this epigenetic transmission of information from one cell generation to the next.

One type of evidence comes from studies in which the nucleus of a cell from a frog or tadpole is transplanted into a frog egg whose own nucleus has been removed (an enucleated egg). In a classic set of experiments performed in 1968, it was shown that a nucleus taken from a differentiated donor cell can be reprogrammed in this way to support development of a whole new tadpole (see Figure 7–2). But this reprogramming occurs only with difficulty, and it becomes less and less efficient as nuclei from older animals are used. Thus, for example, less than 2% of the enucleated eggs injected with a nucleus from a tadpole epithelial cell developed to the swimming tadpole stage, compared with 35% when the donor nuclei were taken instead from an early (gastrula-stage) embryo. With new experimental tools, the cause of this resistance to reprogramming can now be traced. It arises, at least in part, because specific chromatin structures in the original differentiated nucleus tend to persist and be transmitted through the many cell-division cycles required for embryonic development. In experiments with *Xenopus* embryos, specific forms of either repressive or active chromatin structures could be demonstrated to persist through as many as 24 cell divisions, causing the misplaced expression of genes. **Figure 4–45** briefly describes one such experiment,

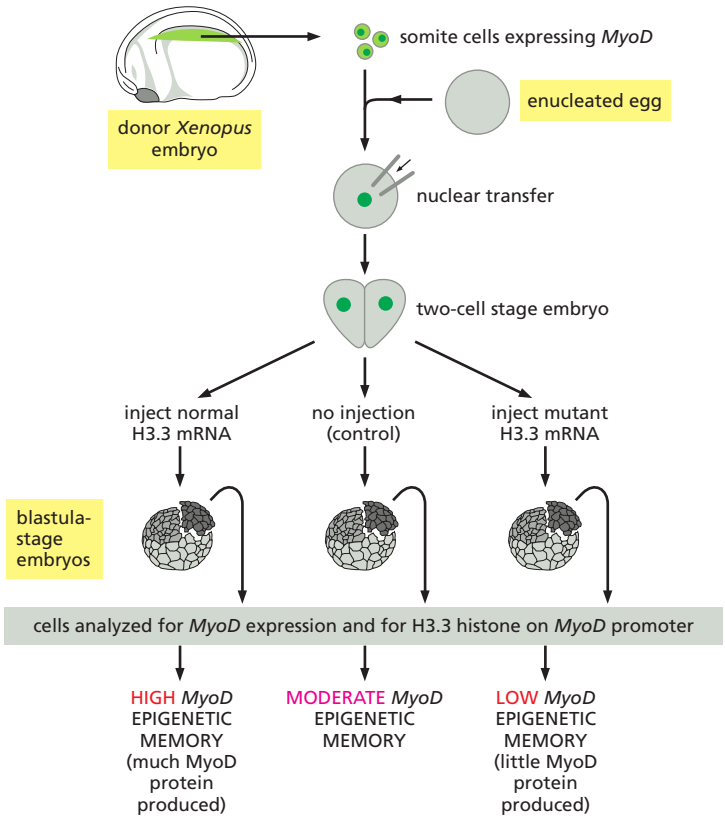


Figure 4–45 Evidence for the inheritance of a gene-activating chromatin state. The well-characterized *MyoD* gene encodes a master transcription regulatory protein for muscle, MyoD (see p. 399). This gene is normally turned on in the indicated region of the young embryo where somites form. When a nucleus from this region is injected into an enucleated egg as shown, many of the progeny cell nuclei abnormally express the MyoD protein in non-muscle regions of the "nuclear transplant embryo" that forms. This abnormal expression can be attributed to maintenance of the *MyoD* promoter region in its active chromatin state through the many cycles of cell division that produce the blastula-stage embryo—a so-called "epigenetic memory" that persists in this case in the absence of transcription. The active chromatin surrounding the *MyoD* promoter contains the variant histone H3.3 (see Figure 4–35) in a Lys4 methylated form. As indicated, an overproduction of this histone caused by injecting excess mRNA encoding the normal H3.3 protein increases both H3.3 occupancy on the *MyoD* promoter and the epigenetic MyoD production, whereas injection of an mRNA producing a mutant form of H3.3 that cannot be methylated at Lys4 reduces the epigenetic MyoD production. Such experiments provide evidence that an inherited chromatin state underlies the epigenetic memory observed. (Adapted from R.K. Ng and J.B. Gurdon, *Nat. Cell Biol.* 10:102–109, 2008. With permission from Macmillan Publishers Ltd.)

focused on chromatin containing the histone variant, H3.3. We shall return to these phenomena in the final section of Chapter 22, where we discuss stem cells and the ways in which one cell type can be converted into another.

Chromatin Structures Are Important for Eukaryotic 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 eukaryotes like ourselves. To form a complex multicellular organism, the cells in different lineages must specialize by changing the accessibility and activity of many hundreds of genes. As described in Chapter 21, this process depends on cell memory: each cell holds a record of its past developmental history in the regulatory circuits that control its many genes. That record, it seems, is partly stored in the structure of the chromatin.

Although bacteria also have cell memory mechanisms, the complexity of the memory circuits in higher eukaryotes is unparalleled. Strategies based on local variations in chromatin structure, unique to eukaryotes, can enable individual genes, once they are switched on or switched off, to stay in that state until some new factor acts to reverse it. At one extreme are structures like centromeric chromatin that, once established, are stably inherited from one cell generation to the next. Likewise, the major "classical" type of heterochromatin, which contains long arrays of the HP1 protein (see Figure 4–39), can persist stably throughout life. In contrast, a form of condensed chromatin that is created by the Polycomb group of proteins serves to silence genes that must be kept inactive in some conditions, but are active in others. The latter mechanism governs the expression of a large number of genes that encode transcription regulators important in early embryonic development, as discussed in Chapter 21. There are many other variant forms of chromatin, some with much shorter lifetimes, often less than the division time of the cell. We shall say more about the variety of chromatin types in the next section.

Summary

In the chromosomes of eukaryotes, DNA is uniformly assembled into nucleosomes, but a variety of different chromatin structures is possible. 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 trimethylation of many different lysine side chains, an important reaction that is incompatible with the acetylation that can occur on the same lysines. Specific combinations of the modifications mark many nucleosomes, governing their interactions with other proteins. These marks are read when protein modules that are part of a larger protein complex bind to the modified nucleosomes in a region of chromatin. These reader proteins then attract additional proteins that perform various functions.

Some reader protein complexes contain a histone-modifying enzyme, such as a histone lysine methylase, that "writes" the same mark that the reader recognizes. A reader–writer–remodeling complex of this type can spread a specific form of chromatin 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 strong evidence for the inheritance of condensed states of chromatin from one cell generation to the next. A similar mechanism appears to be responsible for maintaining the specialized chromatin at centromeres. More generally, the ability to propagate specific chromatin structures across cell generations makes possible an epigenetic cell memory process that plays a role in maintaining the 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 chromatin fiber is made, we now turn to the organization of the chromosome on a more global scale and the way in which its various domains are arranged in space. As a 30-nm fiber, a 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 the molecular details are still largely a mystery, this higher-order packaging almost certainly involves the folding of the chromatin into a series of loops and coils. This chromatin packing is fluid, frequently changing in response to the needs of the cell.

We begin this section by describing some unusual interphase chromosomes that can be easily visualized. Exceptional though they are, these special cases reveal features that are thought to be representative of all interphase chromosomes. Moreover, they provide ways to investigate some fundamental aspects of chromatin structure that we have touched on in the previous section. Next, we describe how a typical interphase chromosome is arranged in the mammalian cell nucleus. Finally, we shall discuss the additional tenfold compaction that chromosomes undergo in the passage from interphase to mitosis.

Chromosomes Are Folded into Large Loops of Chromatin

Insight into the structure of the chromosomes in interphase cells has come from studies of the stiff and enormously extended chromosomes in growing amphibian oocytes (immature eggs). These very unusual **lampbrush chromosomes** (the largest chromosomes known), paired in preparation for meiosis, 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 4–46** and **Figure 4–47**).

In these chromosomes, a given loop always contains the same DNA sequence that 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

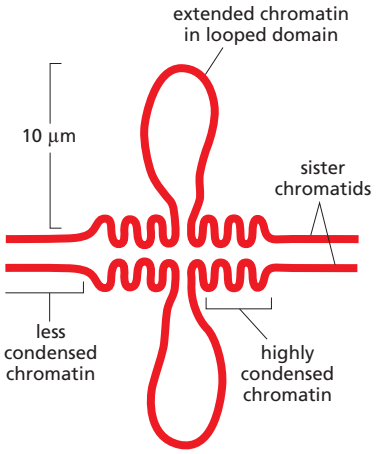


Figure 4–46 A model for the chromatin domains in a lampbrush chromosome. Shown is a small portion of one pair of sister chromatids. Here, two identical DNA double helices are aligned side by side, packaged into different types of chromatin. The set of lampbrush chromosomes in many amphibians contains a total of about 10,000 loops resembling those shown here. The rest of the DNA in each chromosome (the great majority) remains highly condensed. Four copies of each loop are present in the cell, since each lampbrush chromosome consists of two aligned sets of paired chromatids. This four-stranded structure is characteristic of this stage of development of the oocyte, which has arrested at the diplotene stage of meiosis; see Figure 17–56.

present in the DNA loops are being actively expressed. The majority of the DNA, however, is not in loops but remains highly condensed on the chromosome axis, where genes are generally not expressed.

It is thought that the interphase chromosomes of all eukaryotes 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, modern DNA technologies have made it possible to assess the frequency with which any two loci along an interphase chromosome are held together, thus revealing likely candidates for the sites on chromatin that form the bases of loop structures (**Figure 4–48**). These experiments and others suggest that the DNA in human chromosomes is likely to be organized into loops of various 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 4–49**).

Polytene Chromosomes Are Uniquely Useful for Visualizing Chromatin Structures

Further insight has come from another unusual class of cells—the *polytene cells* of flies, such as the fruit fly *Drosophila*. Some types of cells, in many organisms, grow abnormally large through multiple cycles of DNA synthesis without cell division. Such cells, containing increased numbers of standard chromosomes, are said to be *polyploid*. In the salivary glands of fly larvae, this process is taken to an extreme degree, creating huge cells that contain hundreds or thousands of copies of the

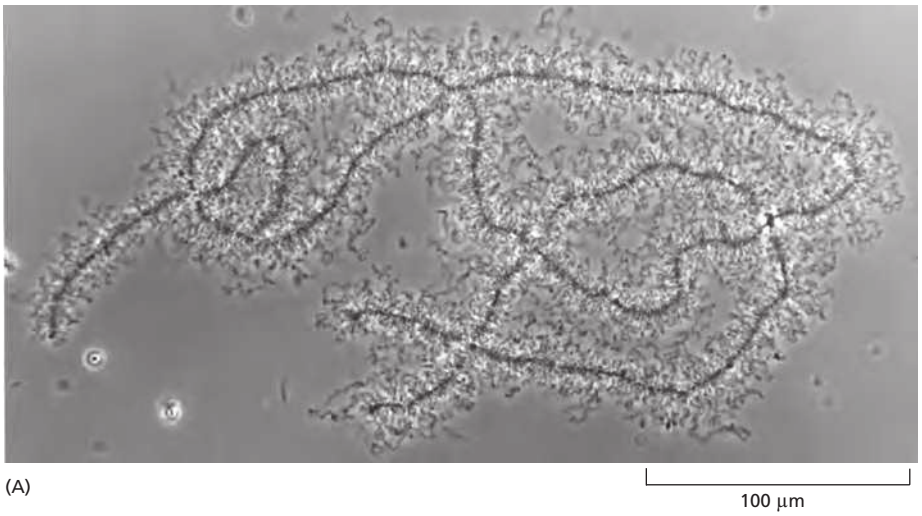
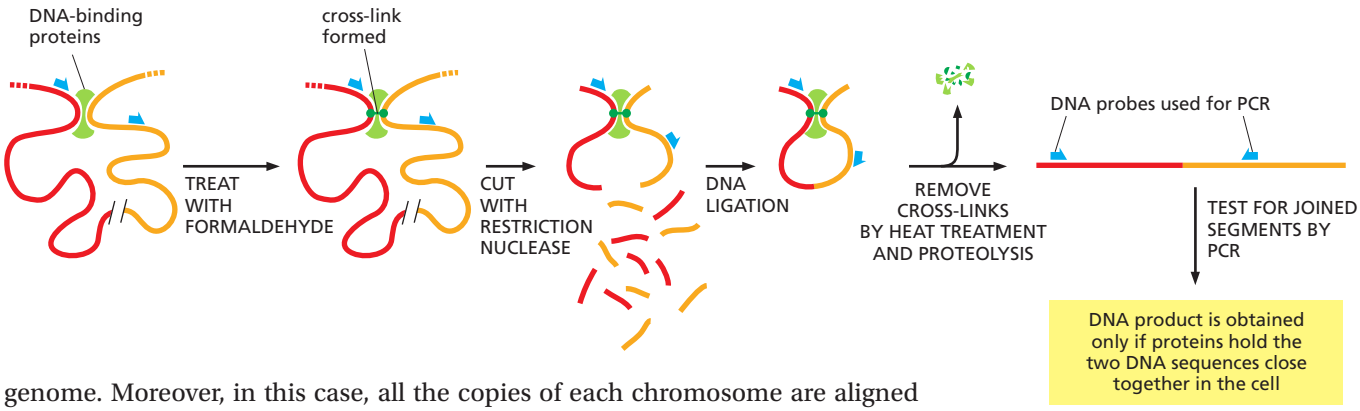


Figure 4–47 Lampbrush chromosomes. (A) A light micrograph of lampbrush chromosomes in an amphibian oocyte. Early in oocyte differentiation, each chromosome replicates to begin meiosis, and the homologous replicated chromosomes pair to form this highly extended structure containing a total of four replicated DNA double helices, 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.)



genome. Moreover, in this case, all the copies of each chromosome are aligned side by side in exact register, like drinking straws in a box, to create giant **polytene chromosomes**. These allow features to be detected that are thought to be shared with ordinary interphase chromosomes, but are normally hard to see.

When polytene chromosomes from a fly's salivary glands are viewed in the light microscope, distinct alternating dark *bands* and light *interbands* are visible (Figure 4–50), 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 concentration of proteins (Figure 4–51). This banding pattern seems to reflect the same sort of organization detected in the amphibian lampbrush chromosomes described earlier.

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. Specific sets of non-histone proteins assemble on these nucleosomes to affect biological function in different ways. Recruitment of some of these non-histone proteins can spread for long distances along the DNA, imparting a similar chromatin structure to broad tracts

Figure 4–48 **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 an enzyme (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 8–28). 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). From the results, combined with DNA sequence information, one can derive models for the interphase conformation of chromosomes.

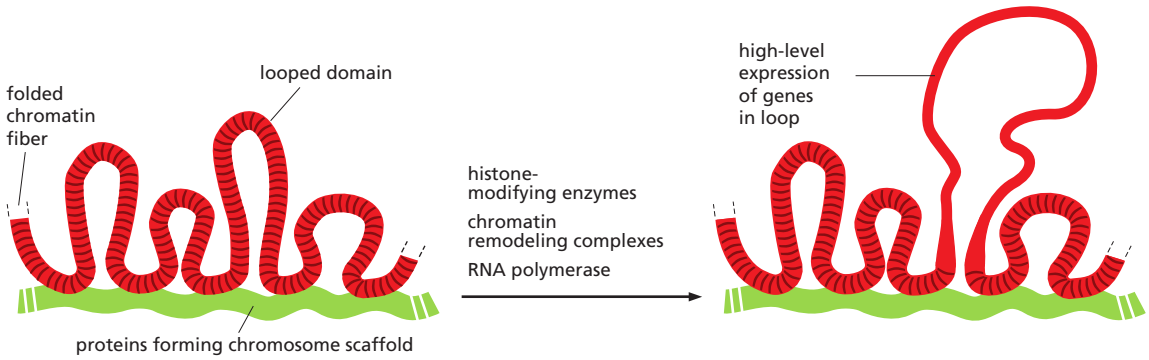


Figure 4–49 **A model for the organization of an interphase chromosome.** A section of an interphase chromosome is shown folded into a series of looped domains, each containing perhaps 50,000–200,000 or more nucleotide pairs of double-helical DNA condensed into a chromatin fiber. The chromatin in each 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 chromatin fiber is anchored to it is clear. However, in mitotic chromosomes, the bases of the chromosomal loops are enriched both in condensins (discussed below) and in DNA topoisomerase II enzymes (discussed in Chapter 5), two proteins that may form much of the axis at metaphase.

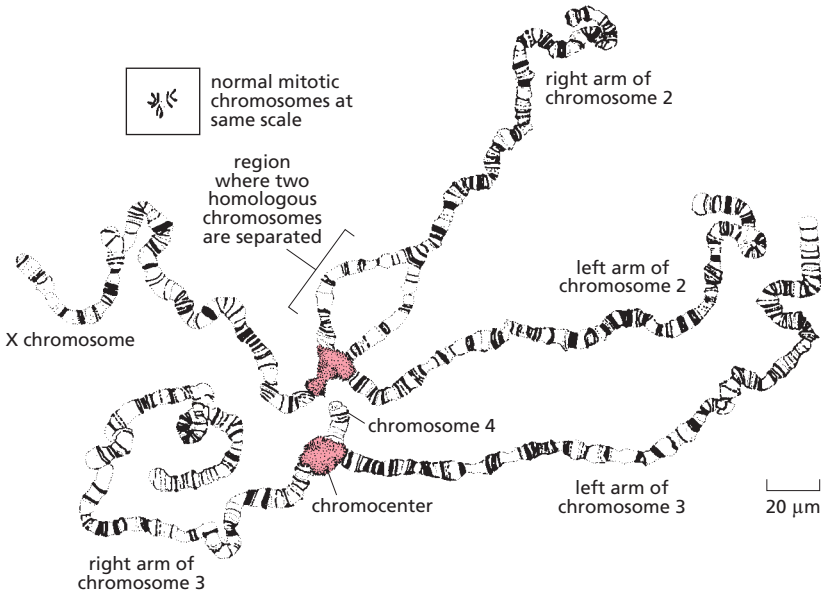


Figure 4–50 The entire set of polytene chromosomes in one *Drosophila* salivary cell. In this drawing of a light micrograph, the giant chromosomes have been spread out for viewing by squashing them against a microscope slide. *Drosophila* has four chromosomes, and there are four different chromosome 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 homologs 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.)

of the genome (see Figure 4–40). Such regions, where all of the chromatin has a similar structure, are separated from neighboring domains by barrier proteins (see Figure 4–41). 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. Polytene chromosomes allow us to see details of this mosaic of domains in the light microscope, as well as to observe some of the changes associated with gene expression.

There Are Multiple Forms of Chromatin

By staining *Drosophila* polytene chromosomes with antibodies, or by using a more recent technique called ChIP (chromatin immunoprecipitation) analysis (see Chapter 8), the locations of the histone and non-histone proteins in chromatin can be mapped across the entire DNA sequence of an organism's genome. Such an analysis in *Drosophila* has thus far localized more than 50 different chromatin proteins and histone modifications. The results suggest that three major types of repressive chromatin predominate in this organism, along with two major types of chromatin on actively transcribed genes, and that each type is associated with a different complex of non-histone proteins. Thus, classical heterochromatin contains more than six such proteins, including heterochromatin protein 1 (HP1),

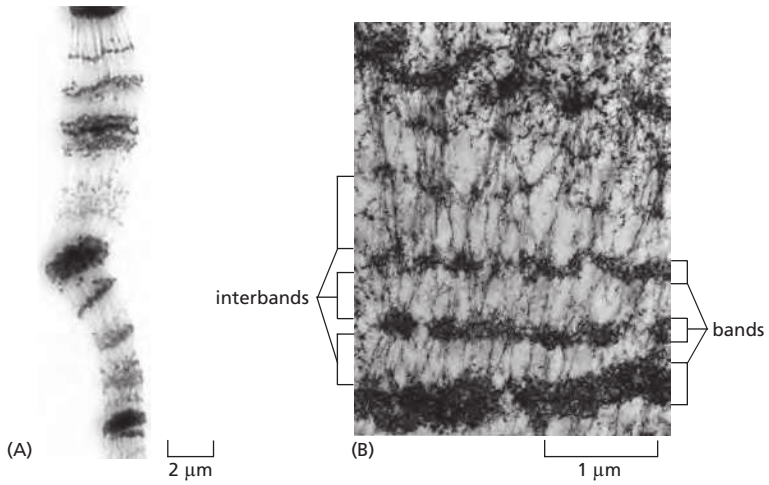


Figure 4–51 Micrographs of polytene chromosomes from *Drosophila* salivary glands. (A) Light micrograph of a portion of a chromosome. The DNA has been stained with a fluorescent dye, but a reverse image is presented here that renders the DNA *black* rather than *white*; the bands are clearly seen to 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.)

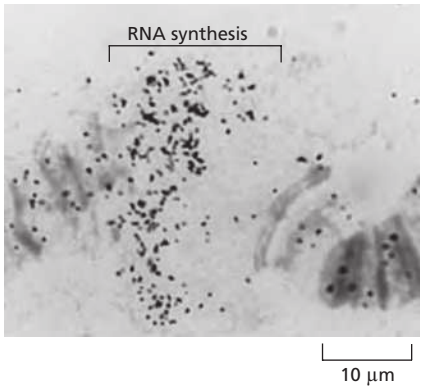


Figure 4–52 RNA synthesis in polytene chromosome puffs. An autoradiograph of a single puff in a polytene chromosome from the salivary glands of the freshwater midge *Chironomus tentans.* As outlined in Chapter 1 and described in detail in Chapter 6, the first step in gene expression is the synthesis 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, an RNA precursor molecule that is incorporated into growing RNA chains. (Courtesy of José Bonner.)

whereas the so-called Polycomb form of heterochromatin contains a similar number of proteins of a different set (PcG proteins). In addition to the five major chromatin types, other more minor forms of chromatin appear to be present, each of which may be differently regulated and have distinct roles in the cell.

The set of proteins bound as part of the chromatin at a given locus varies depending on the cell type and its stage of development. These variations make the accessibility of specific genes different in different tissues, helping to generate the cell diversification that accompanies embryonic development (described in Chapter 21).

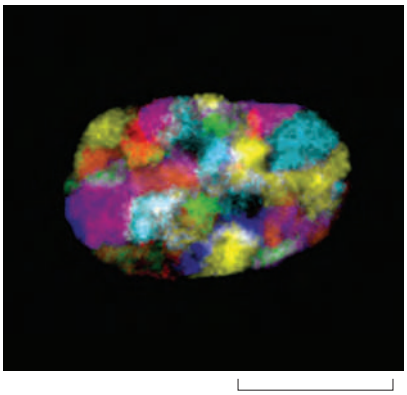
Chromatin Loops Decondense When the Genes Within Them Are Expressed

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 4–52**). 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 amphibian lampbrush chromosomes. When genes in the loop are not expressed, the loop assumes a thickened structure, possibly that of 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 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. However, the same DNA is seen to occupy a larger territory when its genes are expressed, with elongated, punctate structures replacing the original dot.

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: that is, the chromosomes are not extensively entangled with one another (**Figure 4–53**). 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



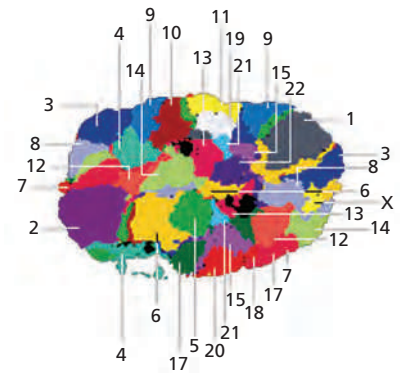


Figure 4–53 Simultaneous visualization of the chromosome territories for all of the human chromosomes in a single interphase nucleus. Here, a mixture of DNA probes for each chromosome has been labeled so as to fluoresce with a different spectra; this allows DNA–DNA hybridization to be used to detect each chromosome, as in Figure 4–10. Three-dimensional reconstructions were then produced. Below the micrograph, each chromosome is identified in a schematic of the actual image. Note that 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.)

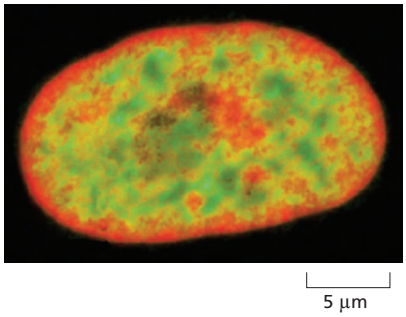


Figure 4–54 The distribution of gene-rich regions of the human genome in an interphase nucleus. Gene-rich regions have been visualized with a fluorescent probe that hybridizes to the *Alu* interspersed repeat, which is present in more than a million copies in the human genome (see page 292). For unknown 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 largely absent in the DNA near the nuclear envelope. (From A. Bolzer et al., *PLoS Biol.* 3:826–842, 2005.)

closely associated with the nuclear lamina, regardless of the chromosome examined. And DNA probes that preferentially stain gene-rich regions of human chromosomes produce a striking picture of the interphase nucleus that presumably reflects different average positions for active and inactive genes (**Figure 4–54**).

How is most of the chromatin in each interphase chromosome condensed when its genes are not being expressed? A powerful extension of the chromosome conformation capture method described previously (see Figure 4–48), which exploits a high-throughput DNA sequencing technology called massive parallel sequencing (see Panel 8–1, pp. 478–481), allows the connections between all of the different one-megabase (1 Mb) segments of the human genome to be mapped in human interphase chromosomes. The results reveal that most regions of our chromosomes are folded into a conformation referred to as a *fractal globule*: a knot-free arrangement that facilitates maximally dense packing while, at the same time, preserving the ability of the chromatin fiber to unfold and fold (**Figure 4–55**).

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

A variety of different types of experiments has 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 sometimes found to extend out of its chromosome territory, as if in an extended loop (**Figure 4–56**). 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 be facilitated in regions of the nucleus enriched 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. It is this issue that we discuss next.

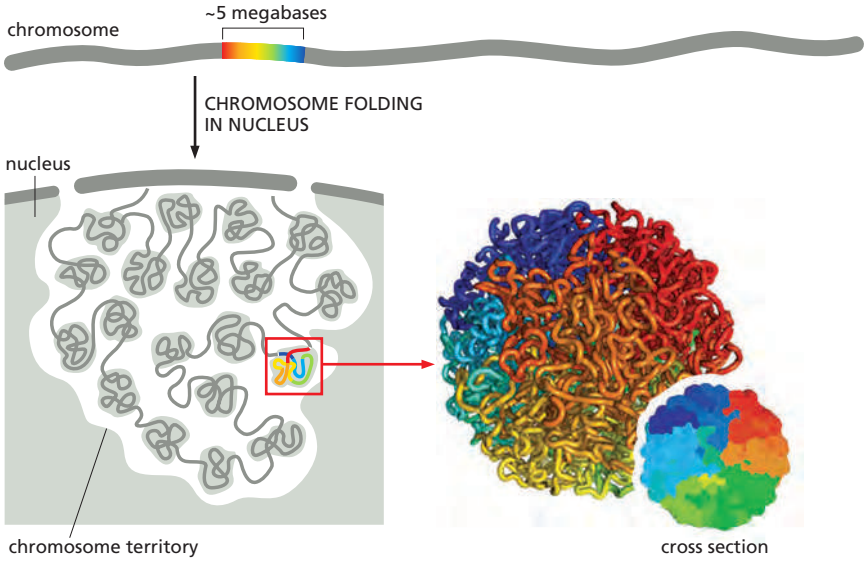


Figure 4–55 A fractal globule model for interphase chromatin. An extension of the 3C method in Figure 4–48, called Hi-C, was used to measure the extent to which each of the three thousand 1 Mb segments in the human genome was located adjacent to any other of these segments. The results support the type of model shown. In the enlarged fractal globule illustrated, a region of 5 million base pairs is seen to fold in a way that keeps regions that are neighbors along the one-dimensional DNA helix as neighbors in three dimensions; this gives rise to monochromatic blocks in this representation that are obvious both on the surface and in cross section. The fractal globule is a knot-free conformation of the DNA that permits dense packing, yet retains an ability to easily fold and unfold any genomic locus. (Adapted from E. Lieberman-Aiden et al., *Science* 326:289–293, 2009. With permission from AAAS.)

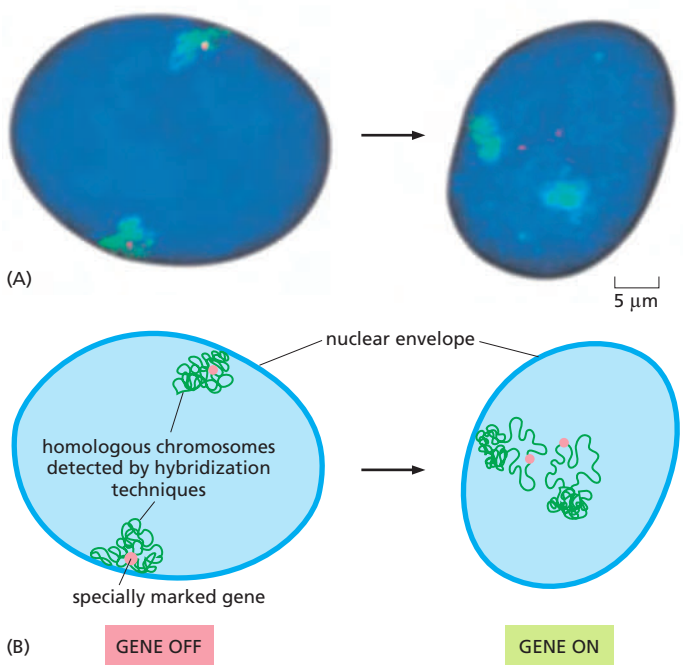


Figure 4–56 An effect of high levels of gene expression on the intranuclear location of chromatin. (A) Fluorescence micrographs of human nuclei showing how the position of a gene changes when it becomes highly transcribed. The region of the chromosome adjacent to the gene *(red)* is seen to leave its 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.)

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

In Chapter 6, we shall 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 nineteenth century (see Figure 4–9). The nucleolus is the cell's site of ribosome subunit formation, as well as the place where many other specialized reactions occur (see Figure 6–42): it consists of a network of RNAs and proteins concentrated around ribosomal RNA genes that are being actively transcribed. In eukaryotes, the genome contains multiple copies of the ribosomal RNA genes, and although they are typically clustered together in a single nucleolus, they are often located on several separate chromosomes.

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 4–57**). 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.

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 other 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. 164) to reactions that take place in subregions of the nucleus (**Figure 4–58**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 environments to perform complex biochemical tasks efficiently. Those that we have mentioned in the nucleus facilitate various aspects of gene expression, and will be further discussed in Chapter 6. These subcompartments, including the nucleolus, appear to form

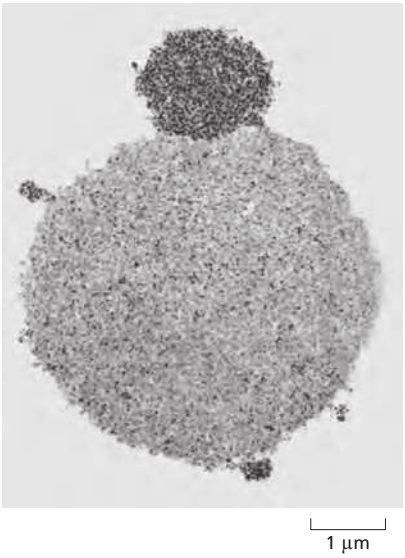


Figure 4–57 Electron micrograph showing two very common fibrous nuclear subcompartments. The large sphere here is a Cajal body. The smaller darker sphere is an interchromatin granule cluster, also known as a speckle (see also Figure 6–46). These "subnuclear organelles" are from the nucleus of a *Xenopus* oocyte. (From K.E. Handwerger and J.G. Gall, *Trends Cell Biol.* 16:19–26, 2006. With permission from Elsevier.)

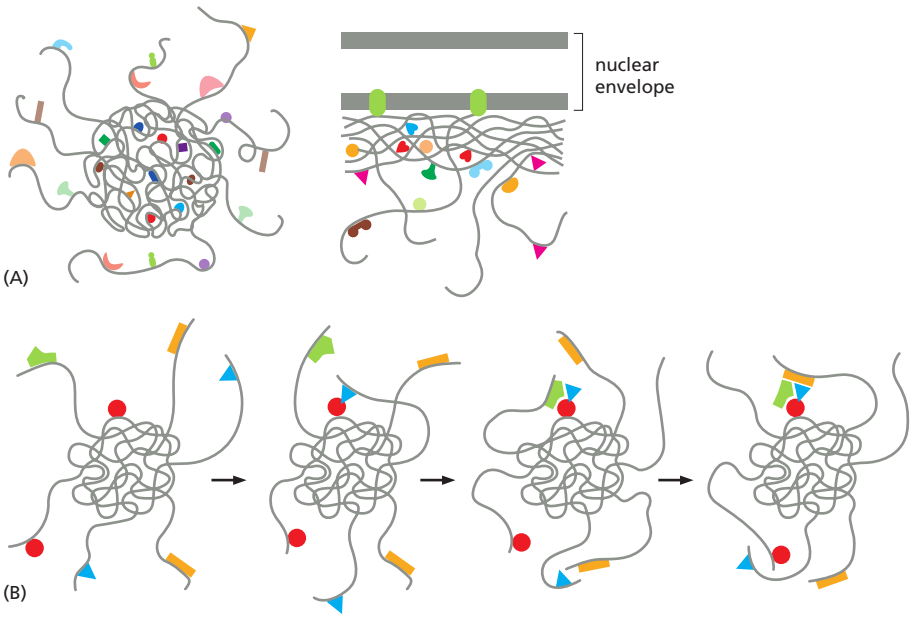


Figure 4–58 Effective compartmentalization without a bilayer membrane. (A) Schematic illustration of the organization of a spherical subnuclear organelle *(left)* and of a postulated similarly organized subcompartment just beneath the nuclear envelope *(right)*. In both cases, RNAs and/or proteins *(gray)* 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), can 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 strategy for accelerating complex sets of reactions is also employed in subcompartments elsewhere in the cell (see also Figure 3–78).

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 5–52). And nuclei often contain hundreds of discrete foci representing factories for DNA or RNA synthesis (see Figure 6–47).

It seems likely that all of these entities make use of the type of tethering illustrated in Figure 4–58B, where long flexible lengths of polypeptide chain and/or long noncoding RNA molecules are interspersed with specific binding sites that concentrate the multiple proteins and other molecules that are needed to catalyze a particular process. Not surprisingly, tethers are similarly used to help to speed biological processes in the cytoplasm, increasing specific reaction rates there (for example, see Figure 16–18).

Is there also 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 may be proteins that help to form the base of chromosomal loops or to attach chromosomes to other structures in the nucleus.

Mitotic Chromosomes Are Especially Highly Condensed

Having discussed the dynamic structure of interphase chromosomes, we now turn to mitotic chromosomes. The chromosomes from nearly all eukaryotic 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 4–59 depicts a typical **mitotic chromosome** at the metaphase stage of mitosis (for the stages of mitosis, see Figure 17–3). The two 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, as mentioned earlier. These chromosomes are normally covered with a variety of molecules, including large amounts of RNA–protein

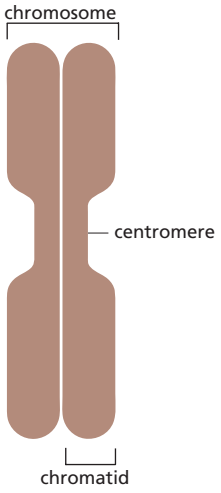


Figure 4–59 A typical mitotic chromosome at metaphase. Each sister chromatid contains one of two identical sister DNA molecules generated earlier in the cell cycle by DNA replication (see also Figure 17–21).

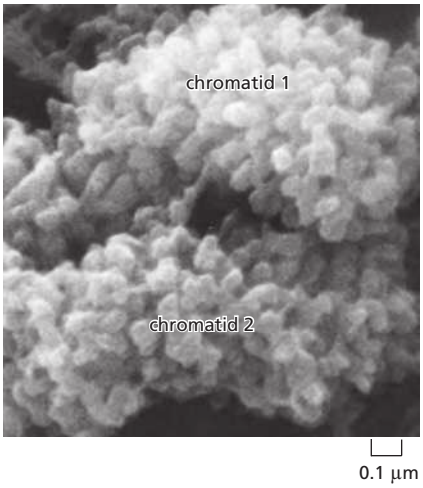


Figure 4–60 A scanning electron micrograph of a region near one end of a typical mitotic chromosome. Each knoblike projection is believed to represent the tip of a separate looped domain. Note that the two identical paired chromatids (drawn in Figure 4–59) can be clearly distinguished. (From M.P. Marsden and U.K. Laemmli, *Cell* 17:849–858, 1979. With permission from Elsevier.)

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 4–60**). 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 4–61**).

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. During M phase, gene expression shuts down, and specific modifications are made to histones that help to reorganize the chromatin as it compacts. Two classes of ring-shaped proteins, called *cohesins* and *condensins*, aid this compaction. How they help to produce the two separately folded chromatids of a mitotic chromosome will be discussed in Chapter 17, along with the details of the cell cycle.

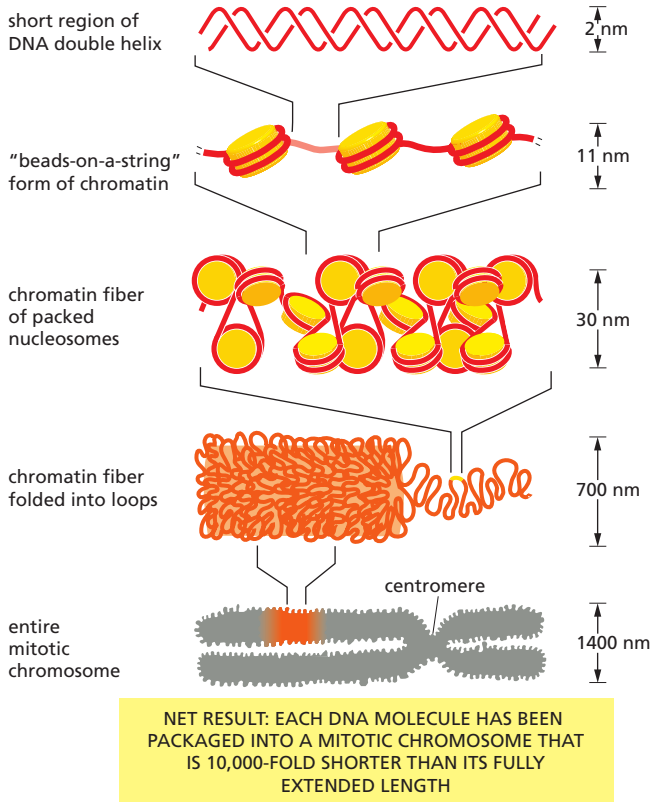


Figure 4–61 Chromatin packing. This model shows some of the many levels of chromatin packing postulated to give rise to the highly condensed mitotic chromosome.

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 that are 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 fibers of compacted nucleosomes. However, euchromatin is interrupted by stretches of heterochromatin, in which the nucleosomes are subjected to additional packing that usually renders the DNA 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 help 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. The condensation is accompanied by histone modifications that facilitate chromatin packing, but 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 additional proteins.

HOW GENOMES EVOLVE

In this final section of the chapter, 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. The sequencing of the genomes of thousands of organisms is revolutionizing our view of the process of evolution, uncovering an astonishing wealth of information about not only family relationships among organisms, but also about the molecular mechanisms by which evolution has proceeded.

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 30 years has been the extent to which the actual nucleotide sequences of many genes have been conserved. **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. Unmistakable homologs of many human genes are present in organisms as diverse as nematode worms, fruit flies, yeasts, and even bacteria. In many cases, the resemblance is so close that, for example, the protein-coding portion of a yeast gene can be substituted with its human homolog—even though humans 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 a sequence match does not guarantee similarity in function, it has proved 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 that have been characterized in other more readily studied organisms.

In general, the sequences of individual genes are much more tightly conserved than is overall genome structure. 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 when comparing distant organisms, as does the number of genes that each organism contains.

Genome Comparisons Reveal Functional DNA Sequences by their Conservation Throughout Evolution

A first obstacle in interpreting the sequence of the 3.2 billion nucleotide pairs in the human genome is the fact that much of it is probably functionally unimportant. The regions of the genome that code for the amino acid sequences of proteins (the exons) are typically found in short segments (average size about 145 nucleotide pairs), small islands in a sea of DNA whose exact nucleotide sequence is thought to be mostly of little consequence. This arrangement makes it difficult to identify all the exons in a stretch of DNA, and it is often hard too to determine exactly where a gene begins and ends.

One very important approach to deciphering our genome is to search for DNA sequences that are closely similar between different species, on the principle that DNA sequences that have a function are much more likely to be conserved than those without a function. For example, 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 pieces of DNA sequence are known as *conserved regions*. In addition to revealing those DNA sequences that encode functionally important exons and RNA molecules, these conserved regions will include regulatory DNA sequences as well as DNA sequences with functions that are not yet known. In contrast, most *nonconserved regions* will reflect DNA whose sequence is much less likely to be critical for function.

The power of this method can be increased by including in such comparisons the genomes of large numbers of species whose genomes have been sequenced, such as rat, chicken, fish, dog, and chimpanzee, as well as mouse and human. 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 our genome. The comparisons reveal that roughly 5% of the human genome consists of "multispecies conserved sequences." To our great surprise, only about one-third of these sequences code for proteins (see Table 4–1, p. 184). Many of the remaining conserved sequences consist of DNA containing clusters of protein-binding sites that are involved in gene regulation, while others produce RNA molecules that are not translated into protein but are important for other known purposes. But, even in the most intensively studied species, the function of the majority of these highly conserved sequences remains unknown. This remarkable discovery has led scientists to conclude that we understand much less about the cell biology of vertebrates than we had thought. Certainly, there are enormous opportunities for new discoveries, and we should expect many more surprises ahead.

Genome Alterations Are Caused by Failures of the Normal Mechanisms for Copying and Maintaining DNA, as well as by Transposable DNA Elements

Evolution depends 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 (discussed below) also plays an important part. As we will explain in Chapter 5, the mechanisms that maintain DNA sequences are remarkably precise—but they are not perfect. DNA sequences are inherited with such extraordinary fidelity that typically, along a given line of descent, only about one nucleotide pair in a thousand is randomly changed in the germ line 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 local changes in DNA sequence—so-called *point mutations* 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, genomes contain mobile DNA elements that are an important source of genomic change (see Table 5–3, p. 267). These transposable DNA elements (*transposons*) are parasitic DNA sequences that can spread within the genomes they colonize. In the process, they often disrupt the function or alter the regulation of existing genes. On occasion, they have created altogether novel genes through fusions between transposon sequences and segments of existing genes. Over long periods of evolutionary time, DNA transposition events have profoundly affected genomes, so much so that nearly half of the DNA in the human genome consists of recognizable relics of past transposition events (**Figure 4–62**). Even more of our genome is thought to have been derived from transpositions 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 Since They Have Separately Evolved

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

The basic organizing framework for comparative genomics is the phylogenetic tree. A simple example is the tree describing the divergence of humans from the great apes (**Figure 4–63**). 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 4–64**). 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

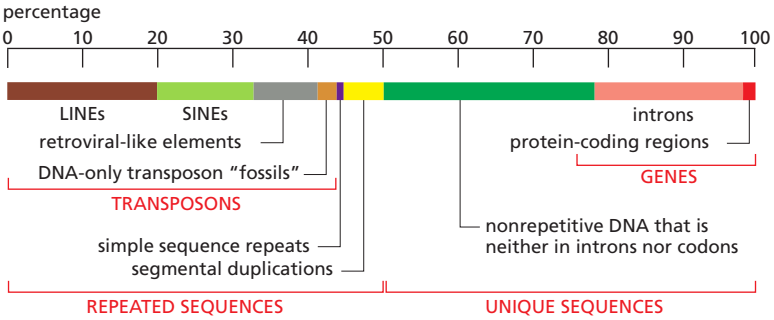


Figure 4–62 A representation of the nucleotide sequence content of the sequenced human genome. The LINEs (long interspersed nuclear elements), SINEs (short interspersed nuclear elements), retroviral-like elements, and DNA-only transposons are mobile genetic elements that have multiplied in our genome 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. 267). 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 DNA sequence (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.)

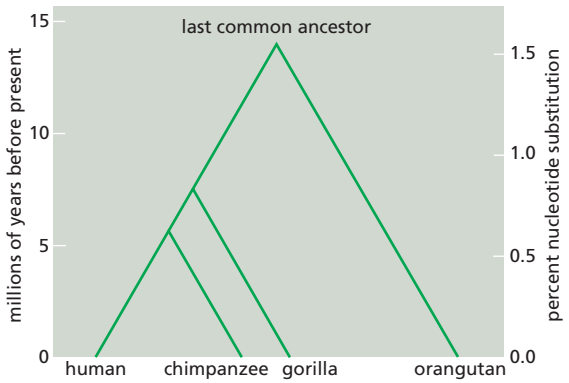


Figure 4–63 A phylogenetic tree showing the relationship between humans and the great apes based on nucleotide sequence data. As indicated, the sequences of the genomes of all four species are estimated to differ from the sequence of the genome of their last common ancestor by a little over 1.5%. Because 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.)

that have kept the sequences the same. Evidence for this view comes from the observation that the human and chimpanzee genomes are nearly identical even where there is no functional constraint on the nucleotide sequence—such as in the third position of "synonymous" codons (codons specifying the same amino acid but differing in their third nucleotide).

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 almost entirely 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.

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

Phylogenetic trees based on molecular sequence data can be compared with the fossil record, and we get our best view of evolution by integrating the two approaches. The fossil record remains essential as a source of absolute dates,

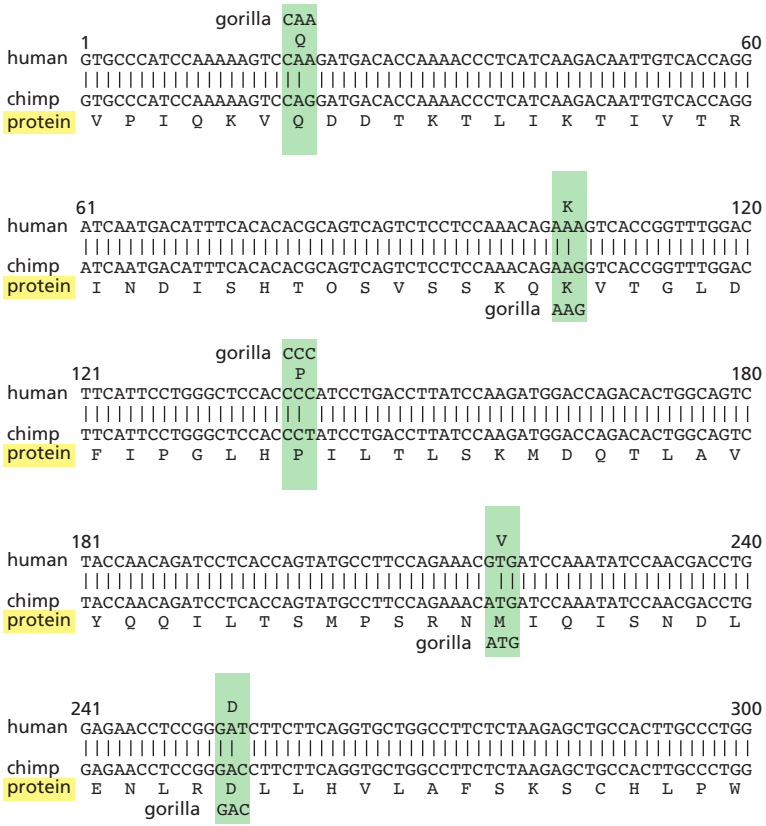


Figure 4–64 Tracing the ancestral sequence from a sequence comparison of the coding regions of human and chimpanzee leptin genes. Reading left to right and top to bottom, a continuous 300-nucleotide segment of a leptin-coding gene is illustrated. Leptin is a hormone that regulates food intake and energy 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 the two species. Moreover, in only one of the five positions does the difference in nucleotide lead to a difference in the encoded amino acid. For each of the five 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? The most economical assumption is that evolution has followed a pathway requiring the minimum number of mutations consistent with the data. Thus, it seems likely that the leptin sequence of the last common ancestor was the same as the human and chimpanzee sequences when they agree; when they disagree, the gorilla sequence would be used as a tiebreaker. For convenience, only the first 300 nucleotides of the leptin-coding sequences are given. The remaining 141 are identical between humans and chimpanzees.

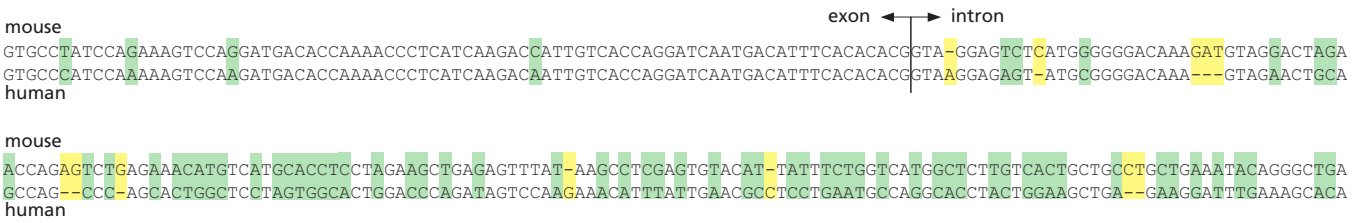


Figure 4–65 The very different rates of evolution of exons and introns, as illustrated by comparing a portion of the mouse and human leptin genes. Positions where the sequences differ by a single nucleotide substitution are boxed in *green*, and positions that differ by the addition or deletion of nucleotides are boxed in *yellow*. Note that, thanks to purifying selection, the coding sequence of the exon is much more conserved than is the adjacent intron sequence.

based on radioisotope decay in the rock formations in which fossils are found. Because the fossil record has many gaps, however, precise divergence times between species are difficult to establish, even for species that leave good fossils with distinctive morphology.

Phylogenetic trees whose timing has been calibrated according to the fossil record 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. This provides us with a *molecular clock* for evolution—or rather a set of molecular clocks corresponding to different categories of DNA sequence. As in the example in **Figure 4–65**, the clock runs most rapidly and regularly in sequences that are not subject to purifying selection. These include portions of introns that lack splicing or regulatory signals, the third position in synonymous codons, 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 that engage in specific interactions with large numbers of other proteins and whose structure is therefore highly constrained, or the nucleotide sequences that encode the RNA subunits of the ribosome, on which all protein synthesis depends.

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 have conferred a selective advantage in the particular lineage where the rapid change occurred.

The pace at which molecular clocks run during evolution is determined not only by the degree of purifying selection, but also by the mutation rate. Most notably, in animals, although not in plants, clocks based on functionally unconstrained mitochondrial DNA sequences run much faster than clocks based on functionally unconstrained nuclear sequences, because the mutation rate in animal mitochondria is exceptionally high.

Categories of DNA for which the clock runs fast are most informative for recent evolutionary events; the mitochondrial DNA clock has been used, for example, to chronicle the divergence of the Neanderthal lineage from that of modern *Homo sapiens*. To study more ancient evolutionary events, one must examine DNA for which the clock runs more slowly; thus the divergence of the major branches of the tree of life—bacteria, archaea, and eukaryotes—has been deduced from study of the sequences specifying ribosomal RNA.

In general, molecular clocks, appropriately chosen, have a finer time resolution than the fossil record, and they are a more reliable guide to the detailed structure of phylogenetic trees than are classical methods of tree construction, which are based on family resemblances in anatomy and embryonic development. For example, the precise family tree of great apes and humans was not settled until sufficient molecular sequence data accumulated in the 1980s to produce the pedigree shown previously in Figure 4–63. And with huge amounts of DNA sequence now determined from a wide variety of mammals, much better estimates of our relationship to them are being obtained (**Figure 4–66**).

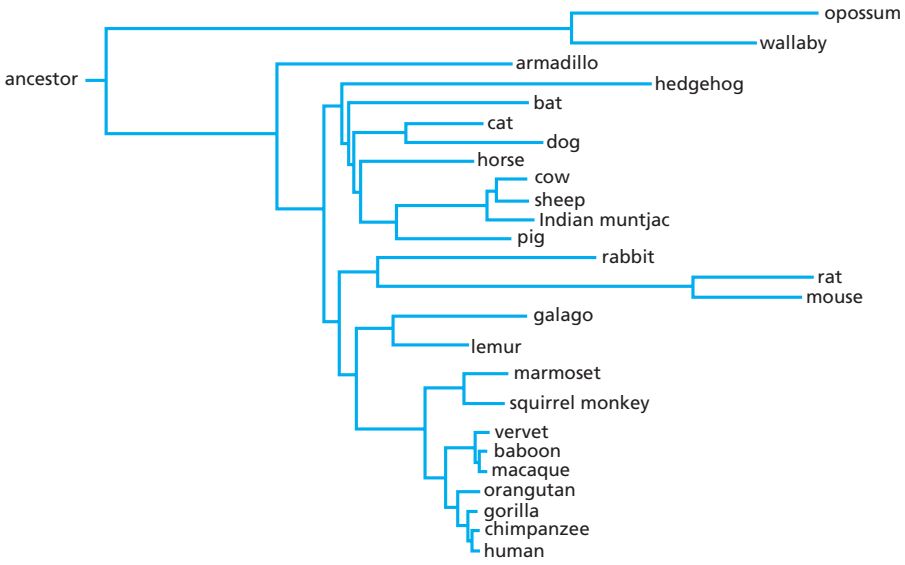


Figure 4–66 A phylogenetic tree showing the evolutionary relationships of some present-day mammals. The length of each line is proportional to the number of "neutral substitutions"—that is, nucleotide changes at sites where there is assumed to be no purifying selection. (Adapted from G.M. Cooper et al., *Genome Res.* 15:901–913, 2005. With permission from Cold Spring Harbor Laboratory Press.)

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, even though all three genomes are roughly the same size and contain nearly identical sets of genes. Mouse and human lineages have had approximately 80 million years to diverge through accumulated mutations, versus 6 million years for humans and chimpanzees. In addition, as indicated in Figure 4–66, rodent lineages (represented by the rat and the mouse) have unusually fast molecular clocks, and have diverged from the human lineage more rapidly than otherwise expected.

While the way that the genome is organized into chromosomes is almost identical between humans and chimpanzees, this organization has diverged greatly between humans and mice. According to rough estimates, a total of about 180 breakage-and-rejoining 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*. **Figure 4–67** illustrates how segments of the different mouse chromosomes map onto the human chromosome set. For much more distantly related vertebrates, such as chicken and human, the number of breakage-and-rejoining events has been much greater and the regions of synteny are much shorter; in addition, they are often hard to discern because of the divergence of the DNA sequences that they contain.

An unexpected conclusion from a detailed comparison of the complete mouse and human genome sequences, confirmed by subsequent comparisons between the genomes of other vertebrates, is that small blocks of DNA sequence 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 3.2 billion nucleotide pairs), mice would have lost a total of about 45% of that genome from accumulated deletions during the past 80 million years, while humans would have lost about 25%. 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 thought to be practically unchanged from that of the last common ancestor of humans and mice, while the mouse genome is smaller by only about 0.3 billion nucleotides.

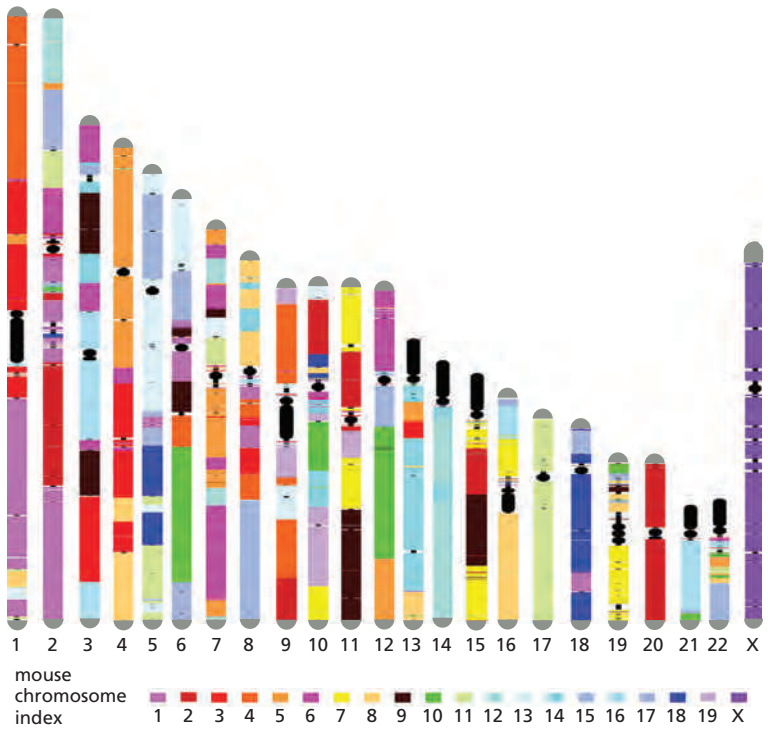


Figure 4–67 Synteny between human and mouse chromosomes. In this diagram, the human chromosome set is shown above, with each part of each chromosome colored according to the mouse chromosome with which it is syntenic. The color coding used for each mouse chromosome is shown below. Heterochromatic highly repetitive regions (such as centromeres) that are difficult to sequence cannot be mapped in this way; these are colored *black*. (Adapted from E.E. Eichler and D. Sankoff, *Science* 301:793–797, 2003. With permission from AAAS.)

Good evidence for the loss of DNA sequences in small blocks during evolution can be obtained from a detailed comparison of 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 4–68**).

DNA is added to genomes both by the spontaneous duplication of chromosomal segments that are typically tens of thousands of nucleotide pairs long (as will be discussed shortly) and by insertion of new copies of active transposons. Most transposition events are duplicative, because the original copy of the transposon stays where it was when a copy inserts at the new site; see, for example, Figure 5–63. Comparison of the DNA sequences derived from transposons in the human and the mouse readily reveals some of the sequence additions (**Figure 4–69**).

It remains a mystery why all mammals have maintained genome sizes of roughly 3 billion nucleotide pairs that contain nearly identical sets of genes, even though only approximately 150 million nucleotide pairs appear to be under sequence-specific functional constraints.

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

In more distantly related vertebrates, genome size can vary considerably, apparently without a drastic effect on the organism or its number of genes. Thus, the chicken genome, at one billion nucleotide pairs, is only about one-third the size

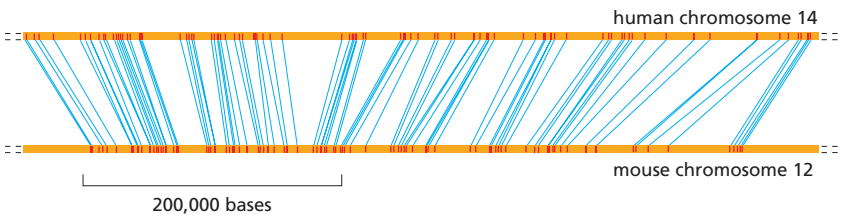


Figure 4–68 Comparison of a syntenic portion of mouse and human genomes. About 90% of the two genomes can be aligned in this way. Note that while there is an identical order of the matched index sequences *(red marks)*, there has been a net loss of DNA in the mouse lineage that is interspersed throughout 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% less DNA than does the human genome. (Adapted from Mouse Genome Sequencing Consortium, *Nature* 420:520–562, 2002. With permission from Macmillan Publishers Ltd.)

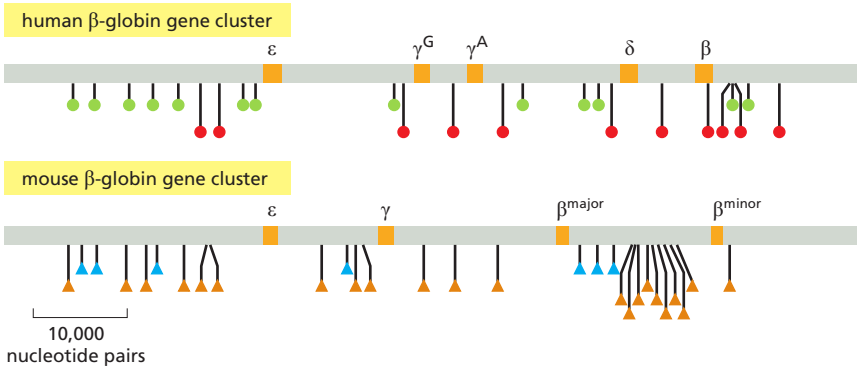


Figure 4–69 A comparison of the β-globin gene cluster in the human and mouse genomes, showing the locations of transposable elements. This stretch of the human genome contains five functional β-globin-like genes *(orange)*; the comparable region from the mouse genome has only four. The positions of the human *Alu* sequences 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.)

of the mammalian genome. An extreme example is the puffer fish, *Fugu rubripes* (**Figure 4–70**), 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 the *Fugu* introns between the exons of each gene are almost the same as in mammalian genomes (**Figure 4–71**).

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. The result is an unusually compact genome, relatively free of junk and clutter, but retaining through purifying selection the vertebrate DNA sequences that are functionally important. This makes *Fugu*, with its 400 million nucleotide pairs of DNA, a valuable resource for genome research aimed at understanding humans.

We Can Infer the Sequence of Some Ancient Genomes

The genomes of ancestral organisms can be inferred, but most can never be directly observed. DNA is very stable compared with most organic molecules, but it is not perfectly stable, and its progressive degradation, even under the best circumstances, means that it is virtually impossible to extract sequence information from fossils that are more than a million years old. 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 in much the same way as in other evolutionary lineages, and at a similar rate. Selection must have maintained key functional properties of the horseshoe-crab genome to account for the morphological stability of the lineage. However, comparisons between different present-day organisms show that the fraction of the genome subject to purifying selection is small; hence, it is fair to assume that the genome of the modern horseshoe crab, while preserving features critical for function, must differ greatly from that of its extinct ancestors, known to us only through the fossil record.

It is possible to get direct sequence information by examining DNA samples from ancient materials if these are not too old. In recent years, technical advances have allowed DNA sequencing from exceptionally well-preserved bone fragments that date from more than 100,000 years ago. Although any DNA this old will be imperfectly preserved, a sequence of the Neanderthal genome has been reconstructed from many millions of short DNA sequences, revealing—among other things—that our human ancestors interbred with Neanderthals in Europe and

Figure 4–70 The puffer fish, *Fugu rubripes*. (Courtesy of Byrappa Venkatesh.)

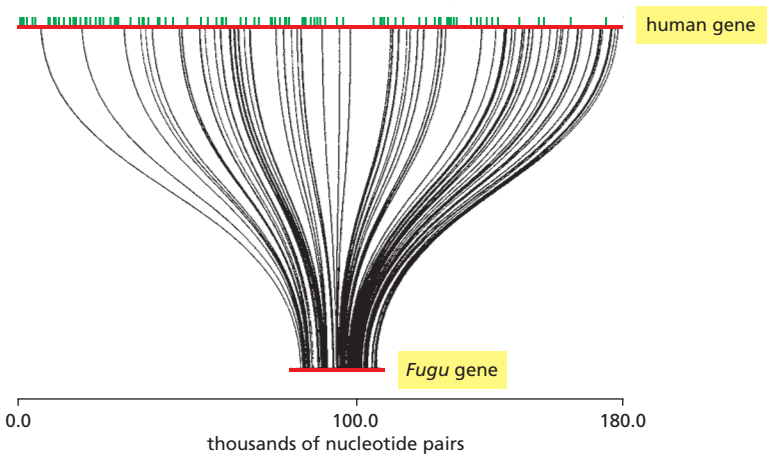


Figure 4–71 Comparison of the genomic sequences of the human and *Fugu* genes encoding the protein huntingtin. Both genes (indicated in *red*) contain 67 short exons that align in 1:1 correspondence to one another; these exons are connected by curved lines. The human gene is 7.5 times larger than the *Fugu* gene (180,000 versus 24,000 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 (discussed in Chapter 5), 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.)

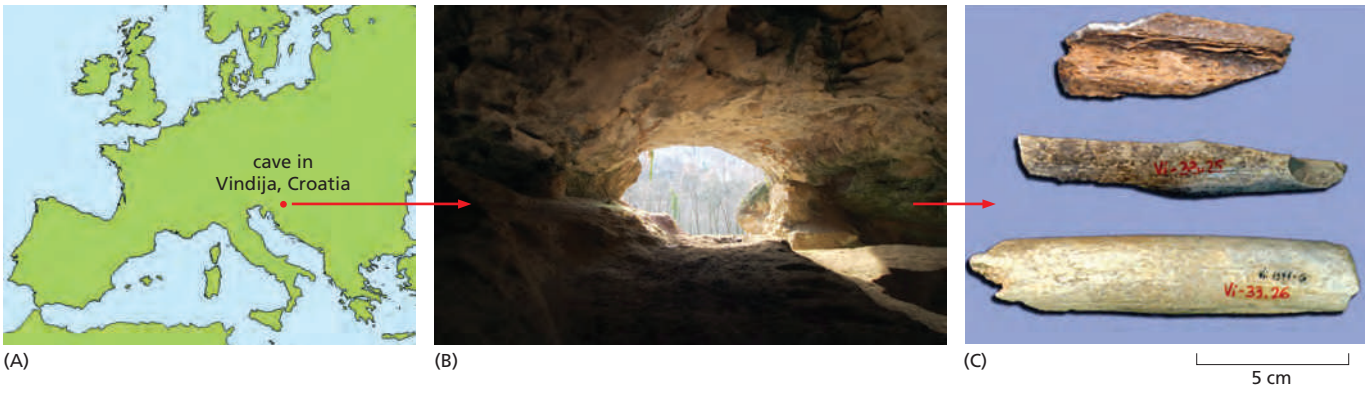
that modern humans have inherited specific genes from them (Figure 4–72). The average difference in DNA sequence between humans and Neanderthals shows that our two lineages diverged somewhere between 270,000 and 440,000 years ago, well before the time that humans are believed to have migrated out of Africa.

But what about deciphering the genomes of much older ancestors, those for which no useful DNA samples can be isolated? For organisms that are as closely related as human and chimpanzee, we saw that this may not be difficult: reference to the gorilla sequence can be used to sort out which of the few sequence differences between human and chimpanzee are inherited from our common ancestor some 6 million years ago (see Figure 4–64). And 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 much of the ancestral sequence, allowing scientists to derive DNA sequences much farther back in time. For example, from the genome sequences currently being obtained for dozens of modern placental mammals, it should be possible to infer much of the genome sequence of their 100 million-year-old common ancestor—the precursor of species as diverse as dog, mouse, rabbit, armadillo, and human (see Figure 4–66).

Multispecies Sequence Comparisons Identify Conserved DNA Sequences of Unknown Function

The mass of DNA sequence now in databases (hundreds of billions of nucleotide pairs) provides a rich resource that scientists can mine for many purposes. This information can be used not only to unscramble the evolutionary pathways that have led to modern organisms, but also to provide insights into how cells and organisms function. Perhaps the most remarkable discovery in this realm comes from the observation that a striking amount of DNA sequence that does not code for protein has been conserved during mammalian evolution (see Table 4–1, p. 184). This is most clearly revealed when we align and compare DNA synteny

Figure 4–72 The Neanderthals. (A) Map of Europe showing the location of the cave in Croatia where most of the bones used to isolate the DNA used to derive the Neanderthal genome sequence were discovered. (B) Photograph of the Vindija cave. (C) Photograph of the 38,000-year-old bones from Vindija. More recent studies have succeeded in extracting DNA sequence information from hominid remains that are considerably older (see Movie 8.3). (B, courtesy of Johannes Krause; C, from R.E. Green et al., *Science* 328: 710–722, 2010. Reprinted with permission from AAAS.)



blocks from many different species, thereby identifying large numbers of so-called *multispecies conserved sequences*: some of these code for protein, but most of them do not (**Figure 4–73**).

Most of the noncoding conserved sequences discovered in this way turn out to be relatively short, containing between 50 and 200 nucleotide pairs. Among the most mysterious are the so-called "ultraconserved" noncoding sequences, exemplified by more than 5000 DNA segments over 100 nucleotides long that are exactly the same in human, mouse, and rat. Most have undergone little or no change since mammalian and bird ancestors diverged about 300 million years ago. The strict conservation implies that even though the sequences do not encode proteins, each nevertheless has an important function maintained by purifying selection. The puzzle is to unravel what those functions are.

Many of the conserved sequences that do not code for protein are now known to produce untranslated RNA molecules, such as the thousands of *long noncoding RNAs* (*lncRNAs*) that are thought to have important functions in regulating gene transcription. As we shall also see in Chapter 7, others are short regions of DNA scattered throughout the genome that directly bind proteins involved in gene regulation. But it is uncertain how much of the conserved noncoding DNA can be accounted for in these ways, and the function of most of it remains a mystery. This enigma highlights how much more we need to learn about the fundamental biological mechanisms that operate in animals and other complex organisms, and its solution is certain to have profound consequences for medicine.

How can cell biologists tackle the mystery of noncoding conserved DNA? Traditionally, attempts to determine the function of a puzzling DNA sequence begin by looking at the consequences of its experimental disruption. But many DNA sequences that are crucial for an organism in the wild can be expected to have no noticeable effect on its phenotype under laboratory conditions: what is required for a mouse to survive in a laboratory cage is very much less than what is required

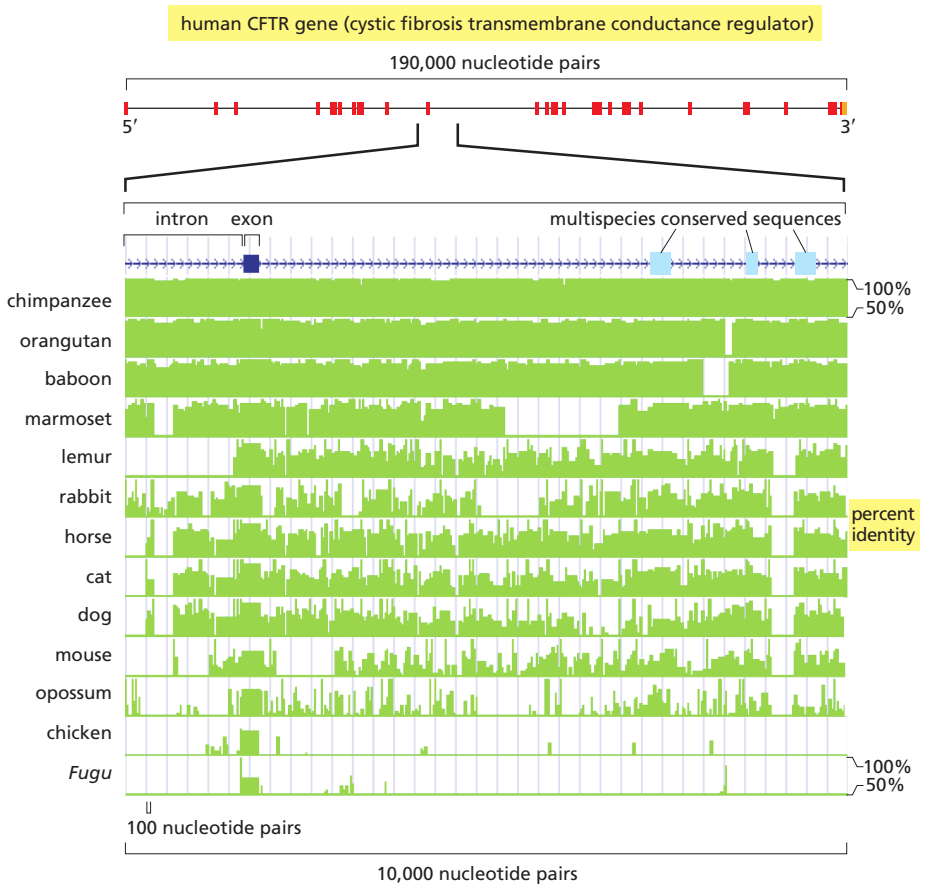


Figure 4–73 The detection of multispecies conserved sequences. In this example, genome sequences for each of the organisms shown have been compared with the indicated region of the human CFTR (cystic fibrosis transmembrane conductance regulator) gene; this region contains one exon plus a large amount of intronic DNA. For each organism, the percent identity with human for each 25-nucleotide block 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 (*dark blue* on the line at the top of the figure), the positions of three other blocks of multispecies conserved sequences are indicated (*pale blue*). The function of most such sequences in the human genome is not known. (Courtesy of Eric D. Green.)

for it to succeed in nature. Moreover, calculations based on population genetics reveal that just a tiny selective advantage—less than a 0.1% difference in survival—can be enough to strongly favor retaining a particular DNA sequence over evolutionary time spans. One should therefore not be surprised to find that many DNA sequences that are ultraconserved can be deleted from the mouse genome without any noticeable effect on that mouse in a laboratory.

A second important approach for discovering the function of a mysterious noncoding DNA sequence uses biochemical techniques to identify proteins or RNA molecules that bind to it—and/or to any RNA molecules that it produces. Most of this task still lies before us, but a start has been made (see p. 435).

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

Given genome sequence information, we can tackle another intriguing question: What alterations in our DNA have made humans so different from other animals—or for that matter, what makes any individual species so different from its relatives? For example, 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 chimpanzees. With 3.2 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 accumulated nucleotide changes exceptionally rapidly during the 6 million years of human evolution since we diverged from the chimpanzees. These *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 chimpanzee, compared to only two changes between chimpanzee and chicken) was examined further and found to encode a 118-nucleotide noncoding RNA molecule, HAR1F (human accelerated region 1F), that is produced in the human cerebral cortex at a critical time during brain development. The function of this HAR1F RNA is not yet known, but findings of this type are stimulating research studies that may shed light on crucial features of the human brain.

A related approach in the search for the important mutations that contributed to human evolution likewise begins with DNA sequences that have been conserved during mammalian evolution, but rather than screening for accelerated changes in individual nucleotides, it focuses instead on chromosome sites that have experienced deletions in the 6 million years since our lineage diverged from that of chimpanzees. More than 500 such sequences—conserved among other species but deleted in humans—have been discovered. Each deletion removes an average of 95 nucleotides of DNA sequence. Only one of these deletions affects a protein-coding region: the rest are thought to alter regions that affect how nearby genes are expressed, an expectation that has been experimentally confirmed in a few cases. A large proportion of the presumed regulatory regions identified in this way lie near genes that affect neural function and/or near genes involved in steroid signaling, suggesting that changes in the nervous system and in immune or reproductive functions have played an especially important role in human evolution.

Mutations in the DNA Sequences That Control Gene Expression Have Driven Many of the Evolutionary Changes in Vertebrates

The vast hoard of genomic sequence data now being accumulated can be explored in many other ways to reveal events that happened even hundreds of millions of years ago. For example, one can attempt to trace the origins of the regulatory elements in DNA that have played critical parts in vertebrate evolution. One such study began with the identification of nearly 3 million noncoding sequences, averaging 28 base pairs in length, that have been conserved in recent vertebrate evolution while being absent in more ancient ancestors. Each of these special non-coding sequences is likely to represent a functional innovation peculiar to a particular branch of the vertebrate family tree, and most of them are thought to consist of regulatory DNA that governs the expression of a neighboring gene. Given full genome sequences, one can identify the genes that lie closest and thus appear most likely to have fallen under the sway of these novel regulatory elements. By comparing many different species, with known divergence times, one can also estimate when each such regulatory element came into existence as a conserved feature. The findings suggest remarkable evolutionary differences between the various functional classes of genes (**Figure 4–74**). Conserved regulatory elements that originated early in vertebrate evolution—that is, more than about 300 million years ago, which is when the mammalian lineage split from the lineage leading to birds and reptiles—seem to be mostly associated with genes that code for transcription regulator proteins and for proteins with roles in organizing embryonic development. Then came an era when the regulatory DNA innovations arose next to genes coding for receptors for extracellular signals. Finally, over the course of the past 100 million years, the regulatory innovations seem to have been concentrated in the neighborhood of genes coding for proteins (such as protein kinases) that function to modify other proteins post-translationally.

Many questions remain to be answered about these phenomena and what they mean. One possible interpretation is that the logic—the circuit diagram—of the gene regulatory network in vertebrates was established early, and that more recent evolutionary change has mainly occurred through the tuning of quantitative parameters. This could help to explain why, among the mammals, for example, the basic body plan—the topology of the tissues and organs—has been largely conserved.

Gene Duplication Also 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 a fish—we rarely encounter genes in the one species that have no homolog in the

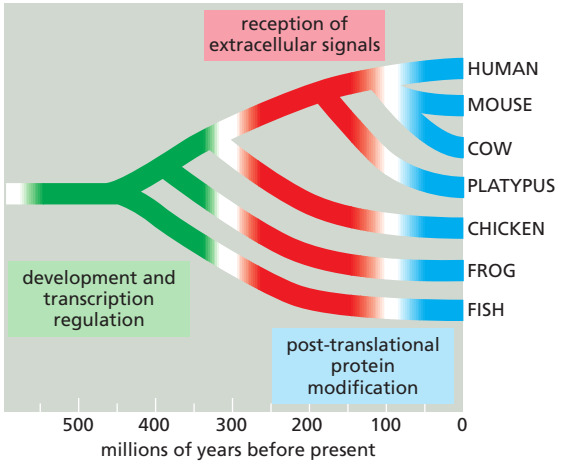


Figure 4–74 The types of changes in gene regulation inferred to have predominated during the evolution of our vertebrate ancestors. To produce the information summarized in this plot, wherever possible the type of gene regulated by each conserved noncoding sequence was inferred from the identity of its closest protein-coding gene. The fixation time for each conserved sequence was then used to derive the conclusions shown. (Based on C.B. Lowe et al., *Science* 333:1019–1024, 2011. With permission from AAAS.)

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.

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, such *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 (although there are some duplications five times larger). In fact, if one counts nucleotides, duplication events have created more differences between our two species than have single-nucleotide substitutions.

Duplicated Genes Diverge

What is 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 equivalent 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 eukaryotic 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 percent 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 germ-line 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 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 occurred in the distant past.

Analysis of the genome of the zebrafish, in which at least one whole-genome duplication is thought to have 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

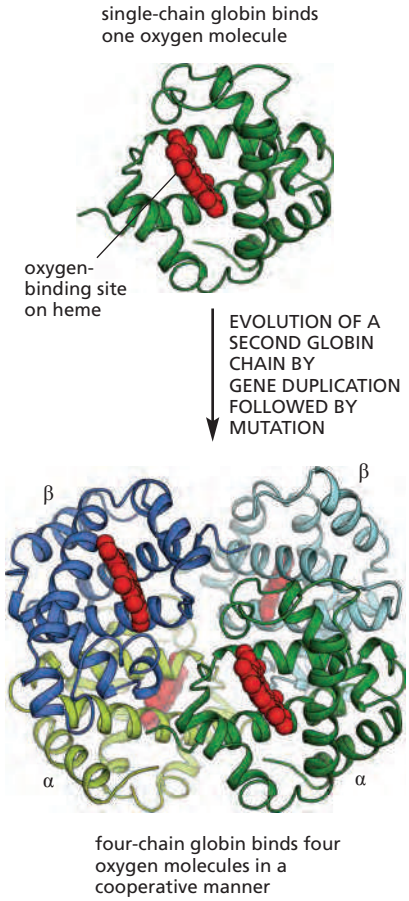


Figure 4–75 A comparison of the structure of one-chain and four-chain globins. The four-chain globin shown is hemoglobin, which is a complex of two α-globin and two β-globin chains. The one-chain globin present in some primitive vertebrates represents an intermediate in the evolution of the four-chain globin. With oxygen bound it exists as a monomer; without oxygen it dimerizes.

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. 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.

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 tree of life. A molecule like hemoglobin was necessary to allow multicellular animals to grow to a large size, since large animals cannot simply rely on the diffusion of oxygen through the body surface to oxygenate their tissues adequately. But oxygen plays a vital part in the life of nearly all living organisms, and oxygen-binding proteins homologous to hemoglobin can be recognized even in plants, fungi, and bacteria. In animals, the most primitive oxygen-carrying molecule is a globin polypeptide chain of about 150 amino acids that 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 in the genome of each individual, coding for α- and β-globin chains that associate to form a hemoglobin molecule consisting of two α chains and two β chains (**Figure 4–75**). 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 4–76**).

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.

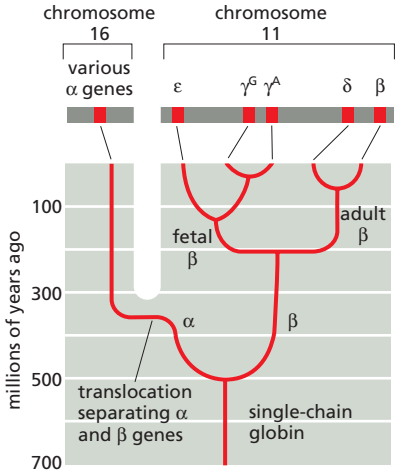


Figure 4–76 An evolutionary scheme for the globin chains that carry oxygen in the blood of animals. The scheme emphasizes the β-like globin gene family. A relatively recent gene duplication of the γ-chain gene produced γ^G and γ^A, which are fetal β-like chains of identical function. The location of the globin genes in the human genome is shown at the top of the figure.

As a result, each globin is made in different amounts at different times of human development.

The history of these gene duplications is reflected in the arrangement of hemoglobin genes in the genome. In the human genome, the genes that arose from the original β gene are arranged as a series of homologous DNA sequences located within 50,000 nucleotide pairs of one another on a single chromosome. A similar cluster of human α-globin genes is located on a separate chromosome. Not only other mammals, but birds too have their α- and β-globin gene clusters on separate chromosomes. In the frog *Xenopus*, however, they are together, suggesting that a chromosome translocation event in the lineage of birds and mammals separated the two gene clusters about 300 million years ago, soon after our ancestors diverged from amphibians (see Figure 4–76).

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 as functional proteins. The existence of such pseudogenes makes it clear that, as expected, not every DNA duplication leads to a new functional gene.

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 4–77**), 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 easily 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 and not disrupt it. 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 complex evolutionary history (see Figure 3–17).

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

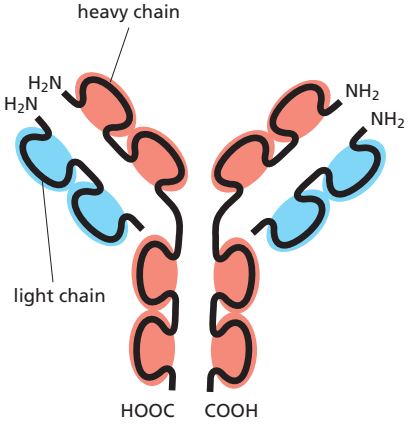


Figure 4–77 Schematic view of an antibody (immunoglobulin) molecule. This molecule is a complex of two identical heavy chains and two identical light chains. Each heavy chain contains four similar, covalently linked domains, while each light chain contains two such domains. Each domain is encoded by a separate exon, and all of the exons are thought to have evolved by the serial duplication of a single ancestral exon.

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 randomly chosen humans, the differences are typically about 0.1%. For more distantly related organisms, the interspecies differences outshine intraspecies 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 for 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 few scattered individuals?

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 evolutionary change in genomes. For example, as we saw earlier, they account for most of the DNA sequence differences between apes and humans. The spread of neutral mutations is not as rapid as the spread of the rare strongly advantageous mutations. It depends on a random variation in the number of mutation-bearing progeny produced by each mutation-bearing individual, causing changes in the relative frequency of the mutant allele in the population. Through a sort of "random walk" process, the mutant allele may eventually become extinct, or it may become commonplace. This can be modeled mathematically for an idealized interbreeding population, on the assumption of 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, study of this idealized case reveals the general principles in a clear and simple way.

When a new neutral mutation occurs in a population of constant size N that is undergoing random mating, the probability that it will ultimately become fixed is approximately $1/(2N)$. This is because there are $2N$ copies of the gene in the diploid population, and each of them has an equal chance of becoming the predominant version in the long run. For those mutations that do become fixed, the mathematics shows that the average time to fixation is approximately $4N$ generations. Detailed analyses of data on human genetic variation have suggested an ancestral population size of approximately 10,000 at the time 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 (1/20,000), 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.

Similar arguments explain another phenomenon with important practical implications for genetic counseling. In an isolated community descended from a small group of founders, such as the people of Iceland or the Jews of Eastern

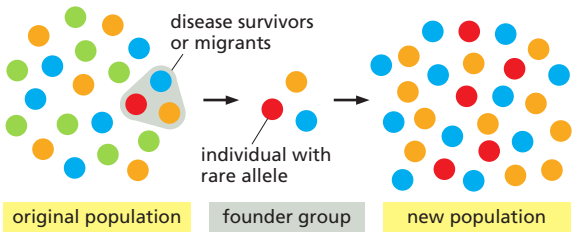


Figure 4–78 How founder effects determine the set of genetic variants in a population of individuals belonging to the same species. This example illustrates how a rare allele *(red)* can become established in an isolated population, even though the mutation that produced it has no selective advantage—or is mildly deleterious.

Europe, genetic variants that are rare in the human population as a whole can often be present at a high frequency, even if those variants are mildly deleterious (**Figure 4–78**).

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

Even though the common variant gene alleles among modern humans originate from variants present in a comparatively tiny group of ancestors, the total number of variants now encountered, including those that are individually rare, is very large. New neutral mutations are constantly occurring and accumulating, even though no single one of them has had enough time to become fixed in the vast modern human population.

From detailed comparisons of the DNA sequences of a large number of modern humans located around the globe, scientists can estimate how many generations have elapsed since the origin of a particular neutral mutation. From such data, it has been possible to map the routes of ancient human migrations. For example, by combining this type of genetic analysis with archaeological findings, scientists have been able to deduce the most probable routes that our ancestors took when they left Africa 60,000 to 80,000 years ago (**Figure 4–79**).

We have been focusing on mutations that affect a single gene, but these are not the only source of variation. Another source, perhaps even more important but missed for many years, lies in the many duplications and deletions of large blocks of human DNA. When one compares any individual human with the standard reference genome in the database, one will generally find roughly 100 differences involving gain or loss of long sequence blocks, totaling perhaps 3 million nucleotide pairs. Some of these **copy number variations** (**CNVs**) will be very common, presumably reflecting relatively ancient origins, while others will be present in only a small minority of people (**Figure 4–80**). On average, nearly half of the CNVs contain known genes. CNVs have been implicated in many human traits, including color blindness, infertility, hypertension, and a wide variety of disease susceptibilities. In retrospect, this type of variation is not surprising, given the prominent role of DNA addition and DNA loss in vertebrate evolution.

The intraspecies variations that have been most extensively characterized, however, 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. To qualify as

Figure 4–79 Tracing the course of human history by analyses of genome sequences. The map shows the routes of the earliest successful human migrations. Dotted lines indicate two alternative routes that our ancestors are thought to have taken out of Africa. DNA sequence comparisons suggest that modern Europeans descended from a small ancestral population that existed about 30,000 to 50,000 years ago. In agreement, archaeological findings suggest that the ancestors of modern native Australians *(solid red arrows)*—and of modern European and Middle Eastern populations—reached their destinations about 45,000 years ago. Even more recent studies, comparing the genome sequences of living humans with those of Neanderthals and another extinct population from southern Siberia (the Denisovans), suggest that our exit from Africa was a bit more convoluted, while also revealing that a number of our ancestors interbred with these hominid neighbors as they made their way across the globe. (Modified from P. Forster and S. Matsumura, *Science* 308:965–966, 2005.)

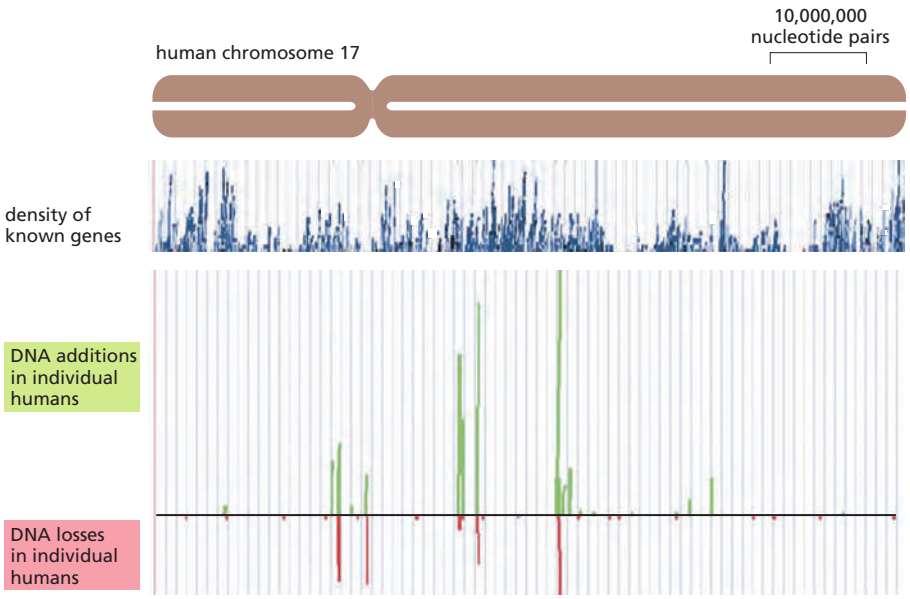


Figure 4–80 Detection of copy number variations on human chromosome 17. When 100 individuals were tested by a DNA microarray analysis that detects the copy number of DNA sequences throughout the entire length of this chromosome, the indicated distributions of DNA additions *(green bars)* and DNA losses *(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.)

a polymorphism, the variants must be common enough to give a reasonably high probability that the genomes of two randomly chosen individuals will differ at the given site; a probability of 1% is commonly chosen as the cutoff. 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, SNPs in the human genome can be extremely useful for genetic mapping analyses, in which one attempts to associate specific traits (phenotypes) with specific DNA sequences for medical or scientific purposes (see p. 493). But while useful as genetic markers, there is good evidence that most of these SNPs have little or no effect on human fitness. This is as expected, since deleterious variants will have been selected against during human evolution and, unlike SNPs, should therefore be rare.

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 eukaryotes. 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, 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 8–39).

While most of the SNPs and CNVs in the human genome sequence are thought to have little or no effect on phenotype, a subset of the genome sequence variations must be responsible for 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 (**Movie 4.3**). We also know that gene dosage—a doubling or halving of the copy number of some genes—can have a profound effect on development by altering the level of gene product, as can changes in regulatory DNA sequences. 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, behavior, and physique. A major challenge in human genetics is to recognize those relatively few variations that are functionally important against a large background of variation that is neutral and of no consequence.

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 a thousand is altered in every million years in any particular eukaryotic 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 (exons)—have been highly conserved. In contrast, nonessential sequences (for example, much of the DNA in introns) have been altered to such an extent that one can no longer see any family resemblance.

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 its functions should teach us many new lessons 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. On a more recent time scale, 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 variations (CNVs). Understanding the effects of these differences will improve both medicine and our understanding of human biology.

WHAT WE DON'T KNOW

- How many different types of chromatin structure are important for cells? How is each of these structures established and maintained, and which ones tend to be inherited following DNA replication?
- Why are there so many different chromatin remodeling complexes in cells? What are their essential roles, and how do they get loaded onto chromatin at specific places and at specific times?
- How do chromosomal loops form during interphase, and what happens to these loops in condensed mitotic chromosomes?
- What genetic changes made us uniquely human? What further aspects of our recent evolutionary development can be reconstructed by sequencing DNA from remains of ancient hominids?
- How much of the enormous complexity that we find in cell biology is unnecessary, having evolved by random drift?

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 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.

4–3 Nucleosomes bind DNA so tightly that they cannot move from the positions where they are first assembled.

4–4 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.

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

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?

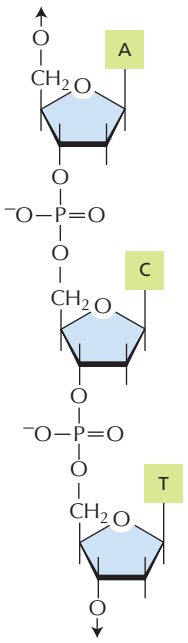


Figure Q4–1 Three nucleotides from the interior of a single strand of DNA (Problem 4–7). *Arrows* at the ends of the DNA strand indicate that the structure continues in both directions.

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 that occurred in the human lineage (**Figure Q4–2**). Draw the intermediate chromosome that resulted from the first inversion and explicitly indicate the segments included in each inversion.

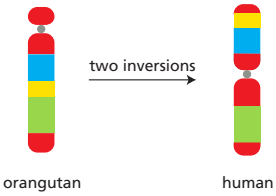


Figure Q4–2 Chromosome 3 in orangutans and humans (Problem 4–9). Differently *colored blocks* indicate segments of the chromosomes that are homologous in DNA sequence.

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 backup 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 Look at the two yeast colonies in **Figure Q4–3**. Each of these colonies contains about 100,000 cells descended from a single yeast cell, originally somewhere in the middle of the clump. A white colony arises when the *Ade2* gene is expressed from its normal chromosomal location. When the *Ade2* gene is moved to a location near a telomere, it is packed into heterochromatin and inactivated in most cells, giving rise to colonies that are mostly red. In these largely red colonies, white sectors fan out from the middle of the colony. In both the red and white sectors, the *Ade2* gene is still located near telomeres. Explain why white sectors have formed near the rim of the red colony. Based on the patterns observed, what can you conclude about the propagation of the transcriptional state of the *Ade2* gene from mother to daughter cells in this experiment?

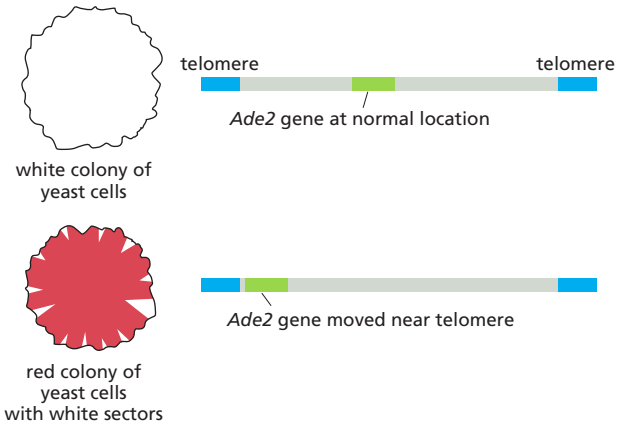


Figure Q4–3 Position effect on expression of the yeast *Ade2* gene (Problem 4–13). The *Ade2* gene codes for one of the enzymes of adenosine biosynthesis, and the absence of the *Ade2* gene product leads to the accumulation of a red pigment. Therefore a colony of cells that express *Ade2* is *white*, and one composed of cells in which the *Ade2* gene is not expressed is *red*.

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 nuclear elements (LINEs), short interspersed nuclear elements (SINEs), long terminal repeat (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?

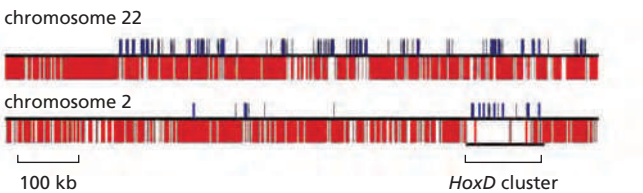


Figure Q4–4 Transposable elements and genes in 1-Mb regions of chromosomes 2 and 22 (Problem 4–14). *Blue lines* that project *upward* indicate exons of known genes. *Red lines* that project *downward* indicate transposable elements; they are so numerous (constituting more than 40% of the human genome) 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

Armstrong L (2014) Epigenetics. New York: Garland Science.

Hartwell L, Hood L, Goldberg ML et al. (2010) Genetics: From Genes to Genomes, 4th ed. Boston, MA: McGraw Hill.

Jobling M, Hollox E, Hurles M et al. (2014) Human Evolutionary Genetics, 2nd ed. New York: Garland Science.

Strachan T & Read AP (2010) Human Molecular Genetics, 4th ed. New York: Garland Science.

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 *Escherichia coli*. *Proc. Natl Acad. Sci. USA* 44, 671–682.

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

Chromosomal DNA and Its Packaging in the Chromatin Fiber

Andrews AJ & Luger K (2011) Nucleosome structure(s) and stability: variations on a theme. *Annu. Rev. Biophys.* 40, 99–117.

Avvakumov N, Nourani A & Côté J (2011) Histone chaperones: modulators of chromatin marks. *Mol. Cell* 41, 502–514.

Deal RB, Henikoff JG & Henikoff S (2010) Genome-wide kinetics of nucleosome turnover determined by metabolic labeling of histones. *Science* 328, 1161–1164.

Grigoryev SA & Woodcock CL (2012) Chromatin organization—the 30 nm fiber. *Exp. Cell Res.* 318, 1448–1455.

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

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

Narlikar GJ, Sundaramoorthy R & Owen-Hughes T (2013) Mechanisms and functions of ATP-dependent chromatin-remodeling enzymes. *Cell* 154, 490–503.

Song F, Chen P, Sun D et al. (2014) Cryo-EM study of the chromatin fiber reveals a double helix twisted by tetranucleosomal units. *Science* 344, 376–380.

Chromatin Structure and Function

Al-Sady B, Madhani HD & Narlikar GJ (2013) Division of labor between the chromodomains of HP1 and Suv39 methylase enables coordination of heterochromatin spread. *Mol. Cell* 51, 80–91.

Beisel C & Paro R (2011) Silencing chromatin: comparing modes and mechanisms. *Nat. Rev. Genet.* 12, 123–135.

Black BE, Jansen LET, Foltz DR & Cleveland DW (2011) Centromere identity, function, and epigenetic propagation across cell divisions. *Cold Spring Harb. Symp. Quant. Biol.* 75, 403–418.

Elgin SCR & Reuter G (2013) Position-effect variegation, heterochromatin formation, and gene silencing in *Drosophila*. *Cold Spring Harb. Perspect. Biol*. 5, a017780.

Felsenfeld G (2014) A brief history of epigenetics. *Cold Spring Harb. Perspect. Biol.* 6, a018200.

Feng S, Jacobsen SE & Reik W (2010) Epigenetic reprogramming in plant and animal development. *Science* 330, 622–627.

Filion GJ, van Bemmel JG, Braunschweig U et al. (2010) Systematic protein location mapping reveals five principal chromatin types in *Drosophila* cells. *Cell* 143, 212–224.

Fodor BD, Shukeir N, Reuter G & Jenuwein T (2010) Mammalian *Su*(*var*) genes in chromatin control. *Annu. Rev. Cell Dev. Biol.* 26, 471–501.

Giles KE, Gowher H, Ghirlando R et al. (2010) Chromatin boundaries, insulators, and long-range interactions in the nucleus. *Cold Spring Harb. Symp. Quant. Biol.* 75, 79–85.

Gohl D, Aoki T, Blanton J et al. (2011) Mechanism of chromosomal boundary action: roadblock, sink, or loop? *Genetics* 187, 731–748.

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

Morris SA, Baek S, Sung M-H et al. (2014) Overlapping chromatin-remodeling systems collaborate genome wide at dynamic chromatin transitions. *Nat. Struct. Mol. Biol.* 21, 73–81.

Politz JCR, Scalzo D & Groudine M (2013) Something silent this way forms: the functional organization of the repressive nuclear compartment. *Annu. Rev. Cell Dev. Biol.* 29, 241–270.

Rothbart SB & Strahl BD (2014) Interpreting the language of histone and DNA modifications. *Biochim. Biophys. Acta* 1839, 627–643.

Weber CM & Henikoff S (2014) Histone variants: dynamic punctuation in transcription. *Genes Dev.* 28, 672–682.

Xu M, Long C, Chen X et al. (2010) Partitioning of histone H3-H4 tetramers during DNA replication-dependent chromatin assembly. *Science* 328, 94–98.

The Global Structure of Chromosomes

Belmont AS (2014) Large-scale chromatin organization: the good, the surprising, and the still perplexing. *Curr. Opin. Cell Biol.* 26, 69–78.

Bickmore W (2013) The spatial organization of the human genome. *Annu. Rev. Genomics Hum. Genet.* 14, 67–84.

Callan HG (1982) Lampbrush chromosomes. *Proc. R. Soc. Lond. B Biol. Sci.* 214, 417–448.

Cheutin T, Bantignies F, Leblanc B & Cavalli G (2010) Chromatin folding: from linear chromosomes to the 4D nucleus. *Cold Spring Harb. Symp. Quant. Biol.* 75, 461–473.

Cremer T & Cremer M (2010) Chromosome territories. *Cold Spring Harb. Perspect. Biol.* 2, a003889.

Lieberman-Aiden E, van Berkum NL, Williams L et al. (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326, 289–293.

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

Moser SC & Swedlow JR (2011) How to be a mitotic chromosome. *Chromosome Res.* 19, 307–319.

Nizami ZF, Deryusheva S & Gall JG (2010) Cajal bodies and histone locus bodies in *Drosophila* and *Xenopus*. *Cold Spring Harb. Symp. Quant. Biol.* 75, 313–320.

Zhimulev IF (1997) Polytene chromosomes, heterochromatin, and position effect variegation. *Adv. Genet.* 37, 1–566.

How Genomes Evolve

Batzer MA & Deininger PL (2002) Alu repeats and human genomic diversity. *Nat. Rev. Genet.* 3, 370–379.

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

Green RE, Krause J, Briggs AW et al. (2010) A draft sequence of the Neandertal genome. *Science* 328, 710–722.

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

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

Kellis M, Wold B, Snyder MP et al. (2014) Defining functional DNA elements in the human genome. *Proc. Natl Acad. Sci. USA* 111, 6131–6138.

Lander ES (2011) Initial impact of the sequencing of the human genome. *Nature* 470, 187–197.

Lee C & Scherer SW (2010) The clinical context of copy number variation in the human genome. *Expert Rev. Mol. Med.* 12, e8.

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.

DNA Replication, Repair, and Recombination

CHAPTER 5

IN THIS CHAPTER

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 illustrate 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 to permit evolutionary adaptation to changing circumstances. We shall see that 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—in particular, those of DNA recombination including the movement within chromosomes of special DNA sequences called transposable elements.

THE MAINTENANCE OF DNA SEQUENCES

Although, as just pointed out, occasional genetic changes enhance the long-term survival of a species through evolution, 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 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 (see Figure 1–24). Under laboratory conditions, *E. coli* divides about once every 30 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 by growing the cells in the presence of a different

sugar, such as glucose, and testing them subsequently to see how many have lost the ability to survive on a lactose diet. 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) approximately once in about 10^6 bacterial cell generations. Stated differently, bacteria display a mutation rate of about three nucleotide changes per 10^{10} 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 humans. In this case, the complete genomes from a family—parents and offspring—were directly sequenced, and a careful comparison revealed that approximately 70 new single-nucleotide mutations arose in the germ lines of each offspring. Normalized to the size of the human genome, the mutation rate is one nucleotide change per 10^8 nucleotides per human generation. This is a slight underestimate because some mutations will be lethal and will therefore be absent from progeny; however, because relatively little of the human genome carries critical information, this consideration has only a small effect on the true mutation rate. It is estimated that approximately 100 cell divisions occur in the germ line from the time of conception to the time of production of the eggs and sperm that go on to make the next generation. Thus, the human mutation rate, expressed in terms of cell divisions (instead of human generations), is approximately 1 mutation/10^{10} nucleotides/cell division.

Although *E. coli* and humans differ greatly in their modes of reproduction and in their generation times, when the mutation rates of each are normalized to a single round of DNA replication, they are both extremely low and within a factor of three of each other. We shall see later in the chapter that the basic mechanisms that ensure these low rates of mutation have been conserved since the very early history of cells on Earth.

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 depend upon to perhaps 30,000. More than this, and the probability that at least one critical component will suffer a damaging mutation becomes catastrophically high. By an extension of the same argument, a mutation frequency tenfold higher would limit an organism to about 3000 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 animal or plant 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 properly maintain the organized structure of the body. Nucleotide changes in somatic cells can give rise to variant cells, some of which, through "local" natural selection, proliferate rapidly at the expense of the rest of the organism. In an extreme case, the result is the uncontrolled cell proliferation that we know as cancer, a disease that causes (in Europe and North America) more than 20% of human deaths each year. These deaths are due largely to an accumulation of changes in the DNA sequences of somatic cells, as discussed in Chapter 20. 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

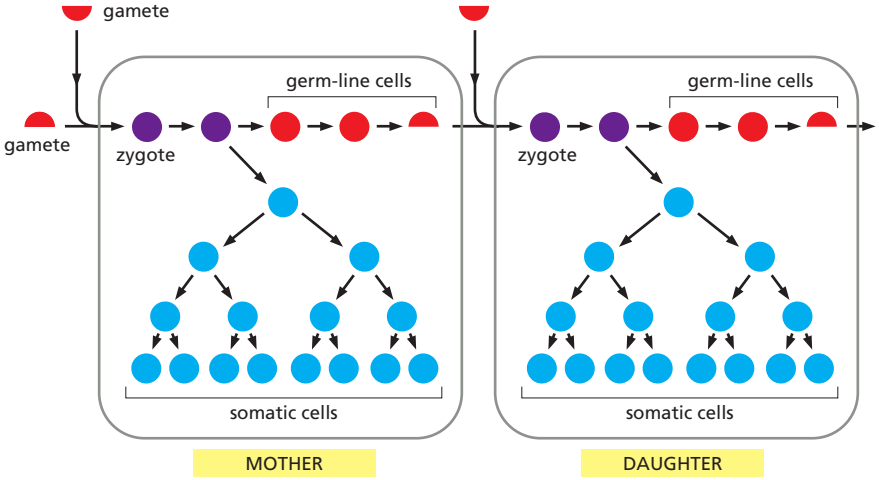


Figure 5–1 Germ-line cells and somatic cells carry out fundamentally different functions. In sexually reproducing organisms, the germ-line cells *(red)* propagate genetic information into the next generation. Somatic cells *(blue)*, which form the body of the organism, are necessary for the survival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

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 one nucleotide change per 10^{10} 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 3.2×10^9 nucleotide pairs) is unchanged or changed by only a few nucleotides each time a typical human 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 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 requires the separation of the DNA helix into two template strands, and entails the recognition of each nucleotide in the DNA *template strands* by a free (unpolymerized) complementary nucleotide. The separation of

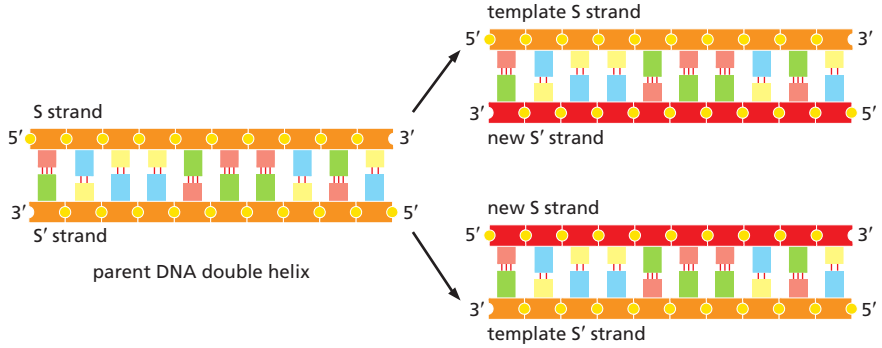


Figure 5–2 The DNA double helix acts as a template for its own duplication. Because the nucleotide A will pair successfully only with T, and G only with C, each strand of DNA can serve as a template to specify the sequence of nucleotides in its complementary strand by DNA base-pairing. In this way, a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

the DNA helix 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-strand 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." How is this feat accomplished?

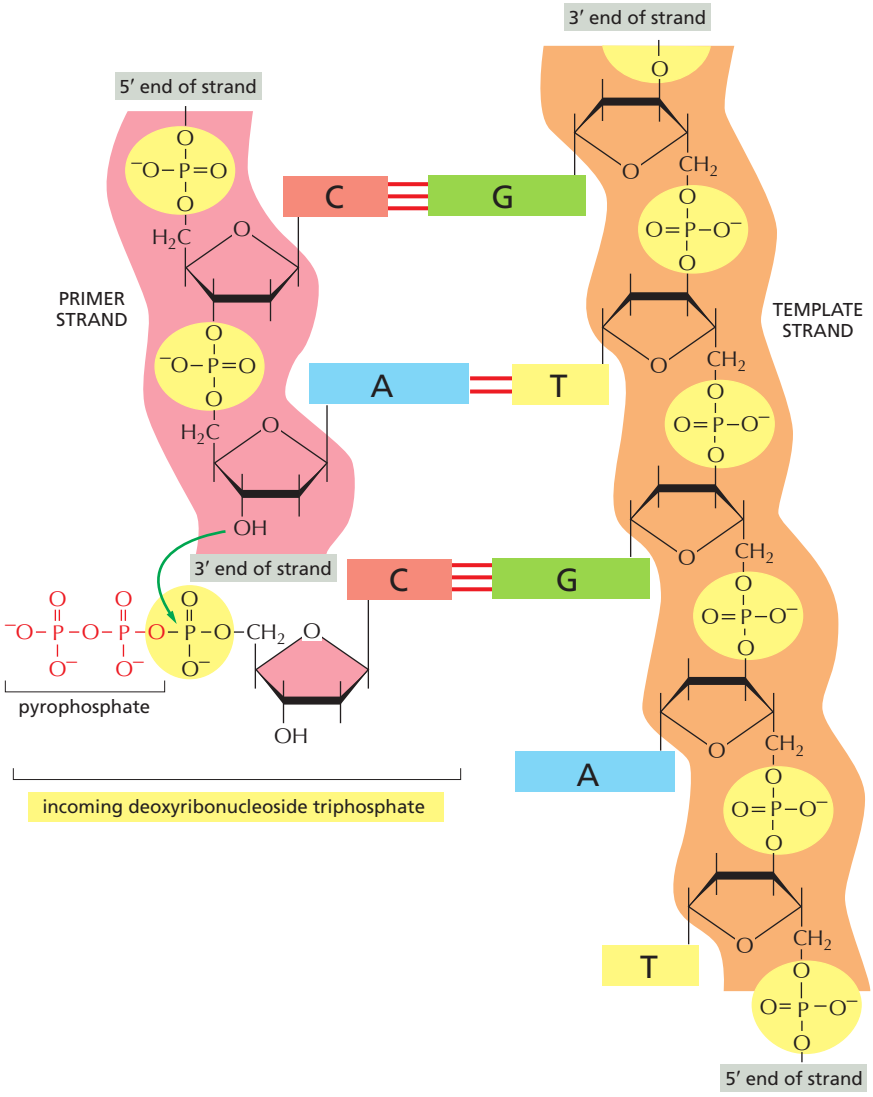


Figure 5–3 The chemistry of DNA synthesis. The addition of a deoxyribonucleotide to the 3′ end of a polynucleotide chain (the *primer strand)* is the fundamental reaction by which DNA is synthesized. As shown, base-pairing between an incoming deoxyribonucleoside triphosphate and an existing strand of DNA (the *template strand)* guides the formation of the new strand of DNA and causes it to have a complementary nucleotide sequence. The way in which complementary nucleotides base-pair is shown in Figure 4–4.

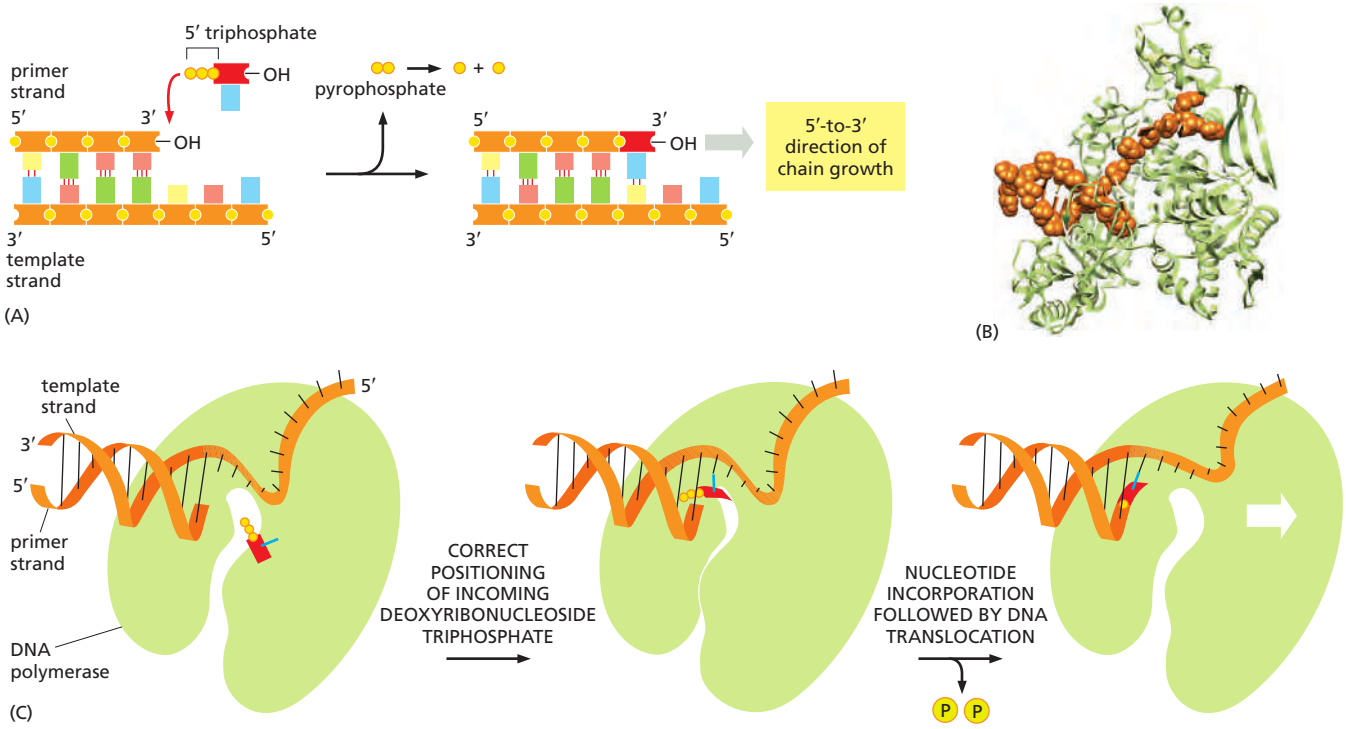


Figure 5–4 DNA synthesis catalyzed by DNA polymerase. (A) DNA polymerase catalyzes the stepwise addition of a deoxyribonucleotide to the 3′-OH end of a polynucleotide chain, the growing *primer strand* that is paired to an existing *template strand*. The newly synthesized DNA strand therefore polymerizes in the 5′-to-3′ direction as shown also 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) Structure of DNA polymerase complexed wth DNA *(orange)*, as determined by x-ray crystallography (**Movie 5.1**). The template DNA strand is the longer strand and the newly synthesized DNA is the shorter. (C) Schematic diagram of DNA polymerase, based on the structure in (B). The proper base-pair geometry of a correct incoming deoxyribonucleoside triphosphate causes the polymerase to tighten around the base pair, thereby initiating the nucleotide addition reaction (*middle diagram* (C)). Dissociation of pyrophosphate relaxes the polymerase, allowing translocation of the DNA by one nucleotide so the active site of the polymerase is ready to receive the next deoxyribonucleoside triphosphate.

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.

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, as attractive as this model might be, the DNA polymerases at replication forks 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 commonly known as *Okazaki fragments,* at the growing replication fork. (Similar replication

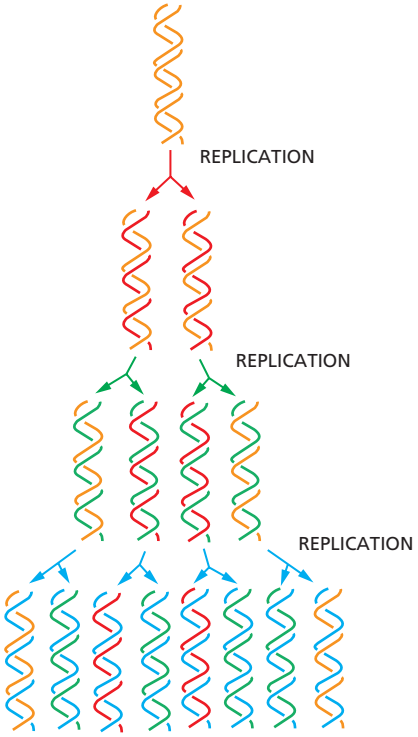


Figure 5–5 The semiconservative nature of DNA replication. In a round of replication, each of the two strands of DNA is used as a template for the formation of a complementary DNA strand. The original strands therefore remain intact through many cell generations.

intermediates were later found in eukaryotes, 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 one mistake occurs for every 10^{10} nucleotides copied. This fidelity is much higher than one would expect from the accuracy of complementary base-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.

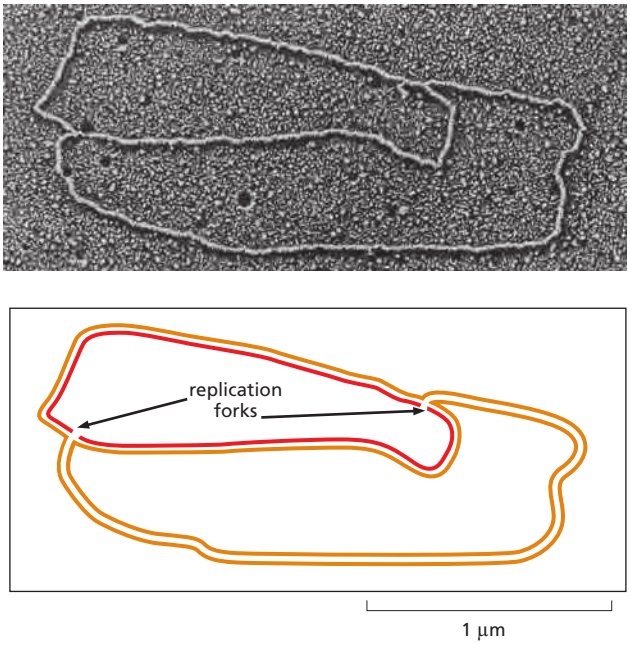


Figure 5–6 Two replication forks moving in opposite directions on a circular chromosome. An active zone of DNA replication moves progressively along a replicating DNA molecule, creating a Y-shaped DNA structure known as a replication fork: the two arms of each Y are the two daughter DNA molecules, and 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.)

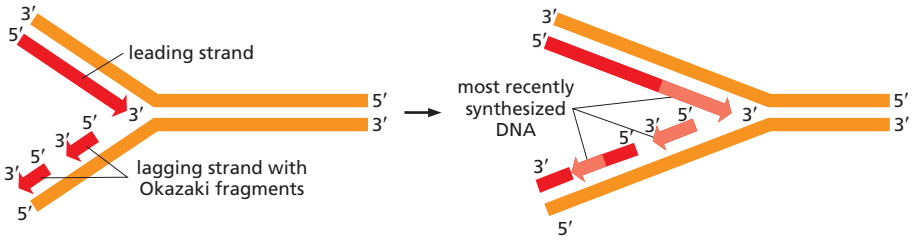


Figure 5–7 The structure of a DNA replication fork. Left, replication fork with newly synthesized DNA in red and arrows indicating the 5′-to-3′ direction of DNA synthesis. Because both daughter DNA strands are polymerized in the 5′-to-3′ direction, the DNA synthesized on the lagging strand must be made initially as a series of short DNA molecules, called *Okazaki fragments*, named after the scientist who discovered them. Right, the same fork a short time later. On the lagging strand, the Okazaki fragments are synthesized sequentially, with those nearest the fork being the most recently made.

DNA polymerase performs the first proofreading step just before a new nucleotide is covalently 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 "grip" tightens 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. Incorrectly paired nucleotides are harder to add and therefore more likely to diffuse away before the polymerase can mistakenly add them.

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 are highly discriminating in the types of DNA chains they will elongate: they 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 has difficulty extending 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 or mispaired 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*, without an existing primer. 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.

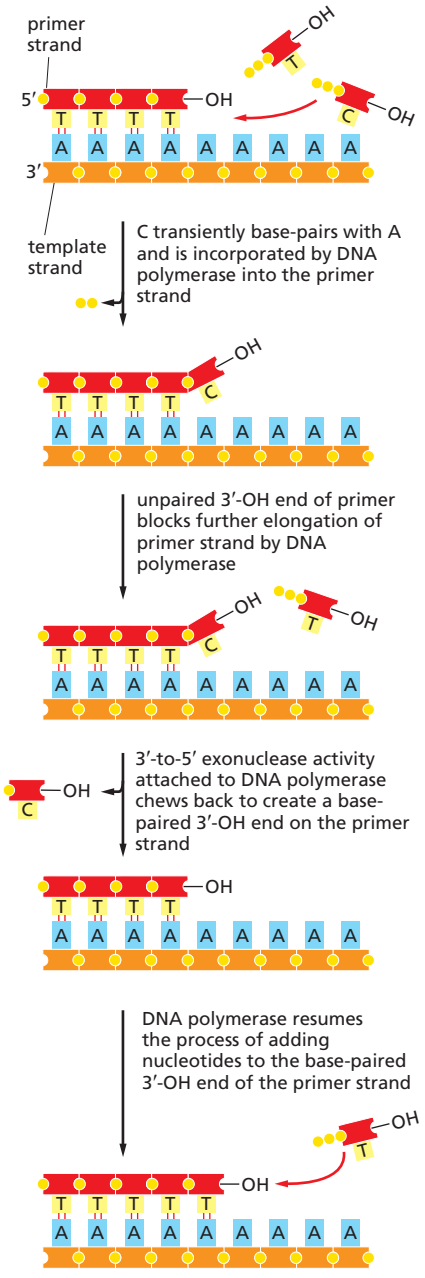


Figure 5–8 Exonucleolytic proofreading by DNA polymerase during DNA replication. In this example, a C is accidentally incorporated at the growing 3′-OH end of a DNA chain. The part of DNA polymerase that removes the misincorporated nucleotide is a specialized member of a large class of enzymes, known as *exonucleases*, that cleave nucleotides one at a time from the ends of polynucleotides.

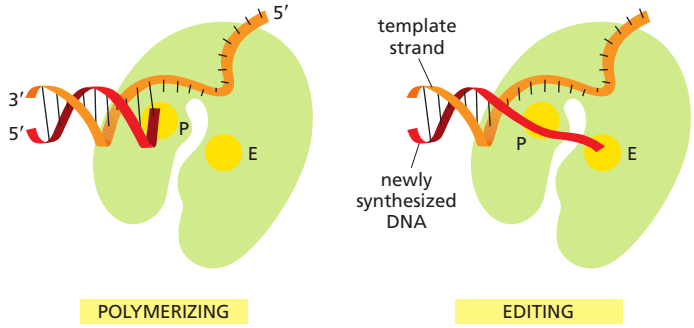


Figure 5–9 Editing by DNA polymerase. DNA polymerase complexed with the DNA template in the polymerizing mode *(left)* and the editing mode *(right)*. The catalytic sites for the exonucleolytic (E) and the polymerization (P) reactions are indicated. In the editing mode, the newly synthesized DNA transiently unpairs from the template and enters the editing site where the most recently added nucleotide is catalytically removed.

There is an error frequency of about one 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 over 100,000 times greater than that in DNA replication, where, as we have seen, a series of proofreading processes makes the process unusually accurate (**Table 5–1**).

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′ end of the chain, rather than the incoming mononucleotide, would have to 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′ end of the chain thus created would immediately terminate DNA synthesis (see Figure 5–3). 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 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

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

Replication step	Errors per nucleotide added
5′ ⟶ 3′ polymerization	1 in 10^5
3′ ⟶ 5′ exonucleolytic proofreading	1 in 10^2
Strand-directed mismatch repair	1 in 10^3
Combined	1 in 10^{10}

The third step, strand-directed mismatch repair, is described later in this chapter. For the polymerization step, "errors per nucleotide added" describes the probability that an incorrect nucleotide will be added to the growing chain. For the other two steps, "errors per nucleotide added" describes the probability that an error will not be corrected. Each step therefore reduces the chance of a final error by the factor shown.

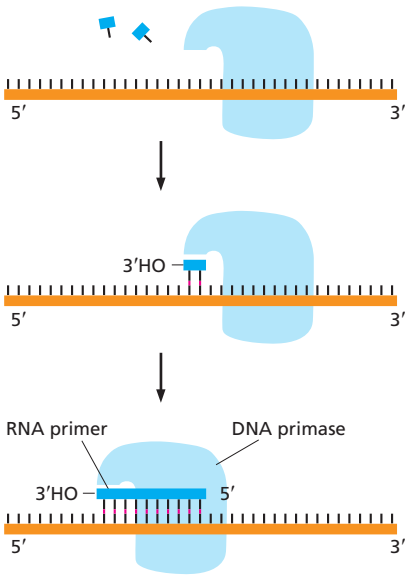


Figure 5–10 RNA primer synthesis. A schematic view of the reaction catalyzed by *DNA primase*, the enzyme that synthesizes the short RNA primers made on the lagging strand using DNA as a template. Unlike DNA polymerase, this enzyme can start a new polynucleotide chain by joining two nucleoside triphosphates 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.

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 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, each 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 the DNA polymerase molecules. The mechanism depends on an enzyme called **DNA primase**, which uses ribonucleoside triphosphates to synthesize short **RNA primers** on the lagging strand (**Figure 5–10**). In eukaryotes, 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 (**Figure 5–11** and **Figure 5–12**).

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 the 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 one 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.

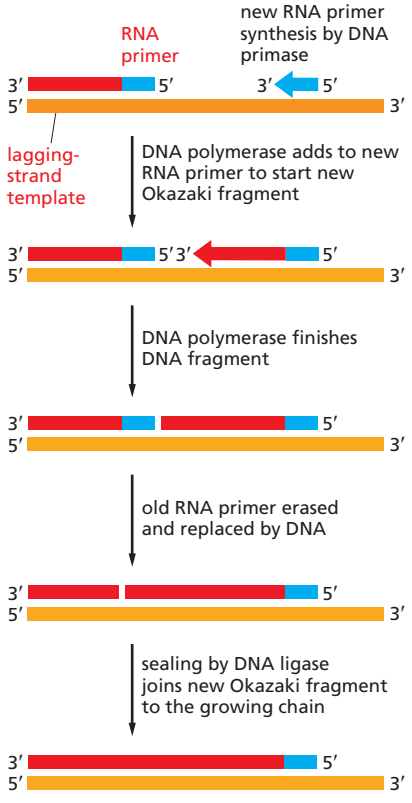


Figure 5–11 The synthesis of one of many DNA fragments on the lagging strand. In eukaryotes, RNA primers are made at intervals spaced by about 200 nucleotides on the lagging strand, and each RNA primer is approximately 10 nucleotides long. This primer is erased by a special DNA repair enzyme (an RNAse 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.

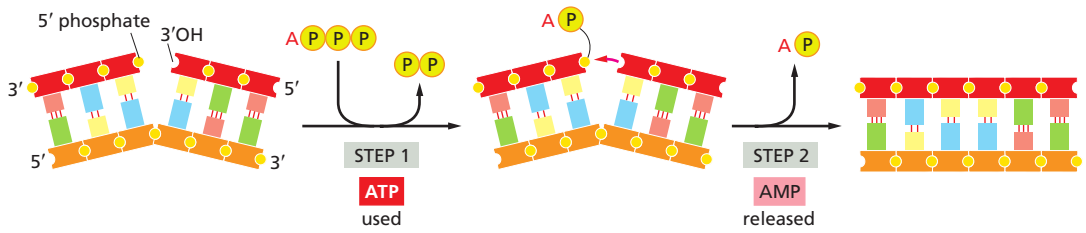


Figure 5–12 The reaction catalyzed by DNA ligase. This enzyme seals a broken phosphodiester bond. As shown, DNA ligase uses a molecule of ATP to activate the 5′ end at the nick (step 1) before forming the new bond (step 2). In this way, the energetically unfavorable nick-sealing reaction is driven by 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 ("melted") ahead of the replication fork so that the incoming deoxyribonucleoside triphosphates can form base pairs with the template strands. 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 provide the appropriate single-strand DNA templates 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 (**Figure 5–13** and **Figure 5–14**).

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-strand DNA without covering the bases, which therefore remain available as templates. These proteins are unable to open a long DNA helix directly, but they aid helicases by stabilizing the unwound, single-strand conformation. In addition, through cooperative binding, they coat and straighten out the regions of single-strand DNA, which occur routinely in the lagging-strand template, thereby preventing the formation of the short hairpin helices that readily form in single-strand DNA (**Figure 5–15** and **Figure 5–16**). If not removed, 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 quickly from a DNA molecule allows a DNA polymerase molecule that has just

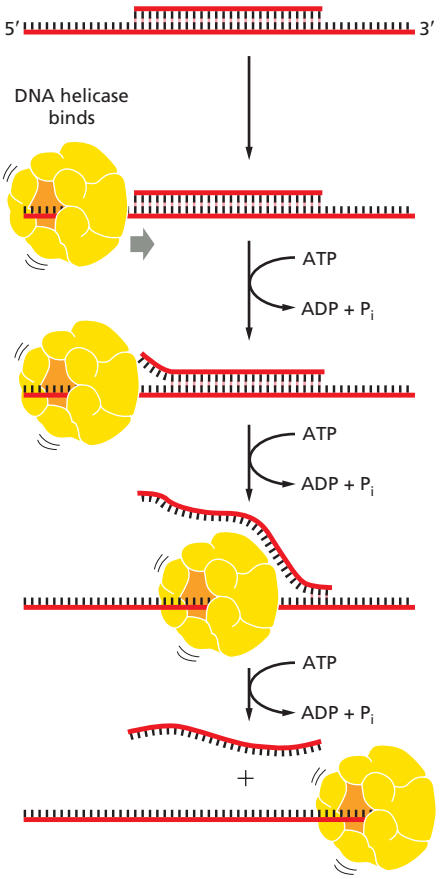
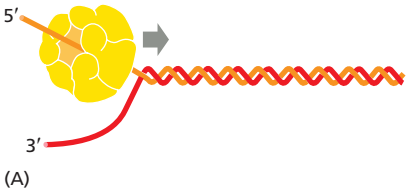
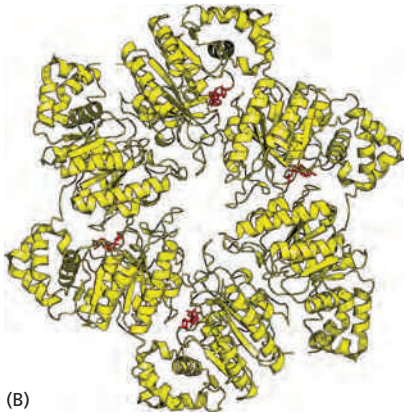


Figure 5–13 An assay for DNA helicase enzymes. A short DNA fragment is annealed to a long DNA single strand to form a region of DNA double helix. The double helix is melted as the helicase runs along the DNA single strand, releasing the short DNA fragment in a reaction that requires the presence of both the helicase protein and ATP. The rapid stepwise movement of the helicase is powered by its ATP hydrolysis (shown schematically in Figure 3–75A). As indicated, many DNA helicases are composed of six subunits.



(A)



(B)

Figure 5–14 The structure of a DNA helicase. (A) Diagram of the protein as a hexameric ring drawn to scale with a replication fork. (B) 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 (Movie 5.2). (PDB code: 1E0J.)

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 (called PCNA in eukaryotes) 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-strand 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, revealed it to be a large ring around the DNA double helix. One face 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 5–17).

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

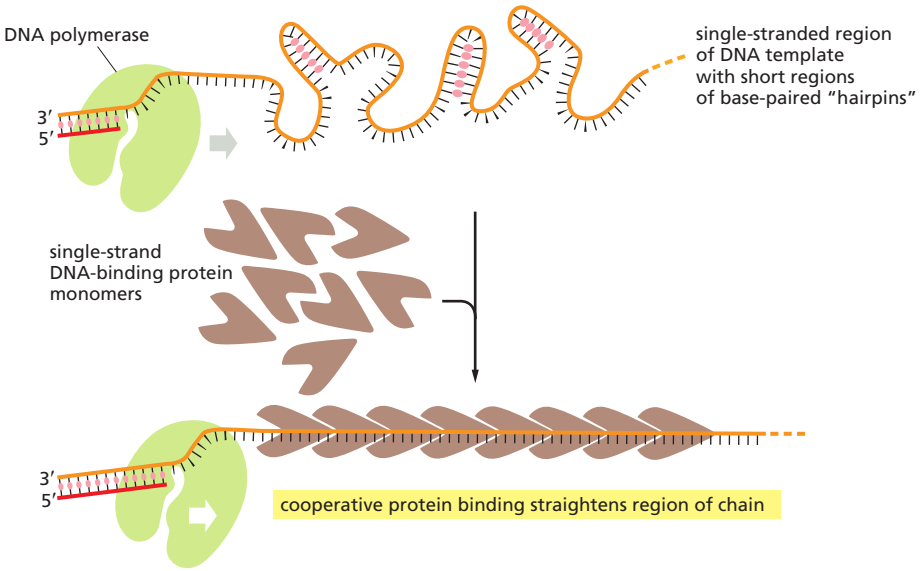


Figure 5–15 The effect of single-strand DNA-binding proteins (SSB proteins) on the structure of single-strand DNA. Because each protein molecule prefers to bind next to a previously bound molecule, long rows of this protein form on a DNA single strand. This *cooperative binding* straightens out the DNA template and facilitates the DNA polymerization process. The "hairpin helices" shown in the bare, single-strand 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).

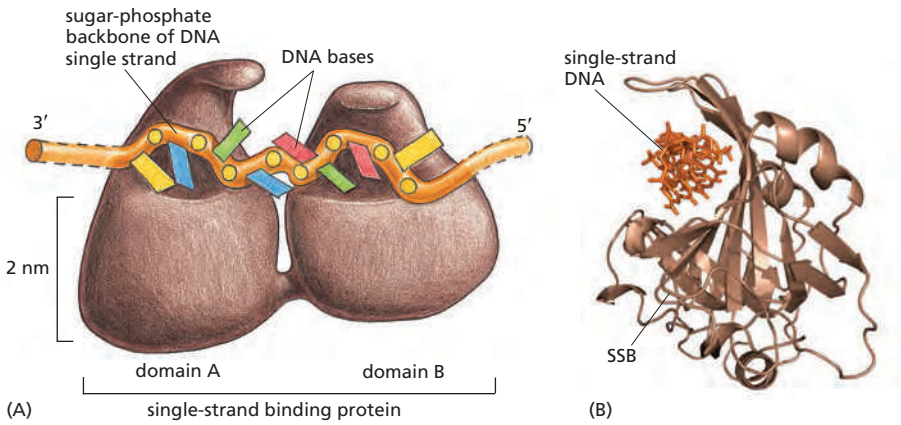
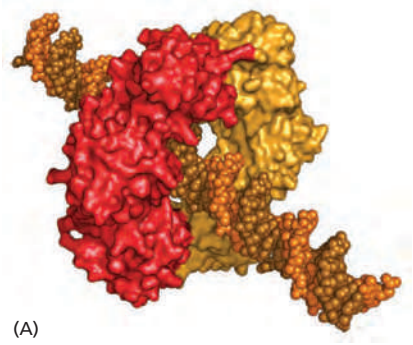
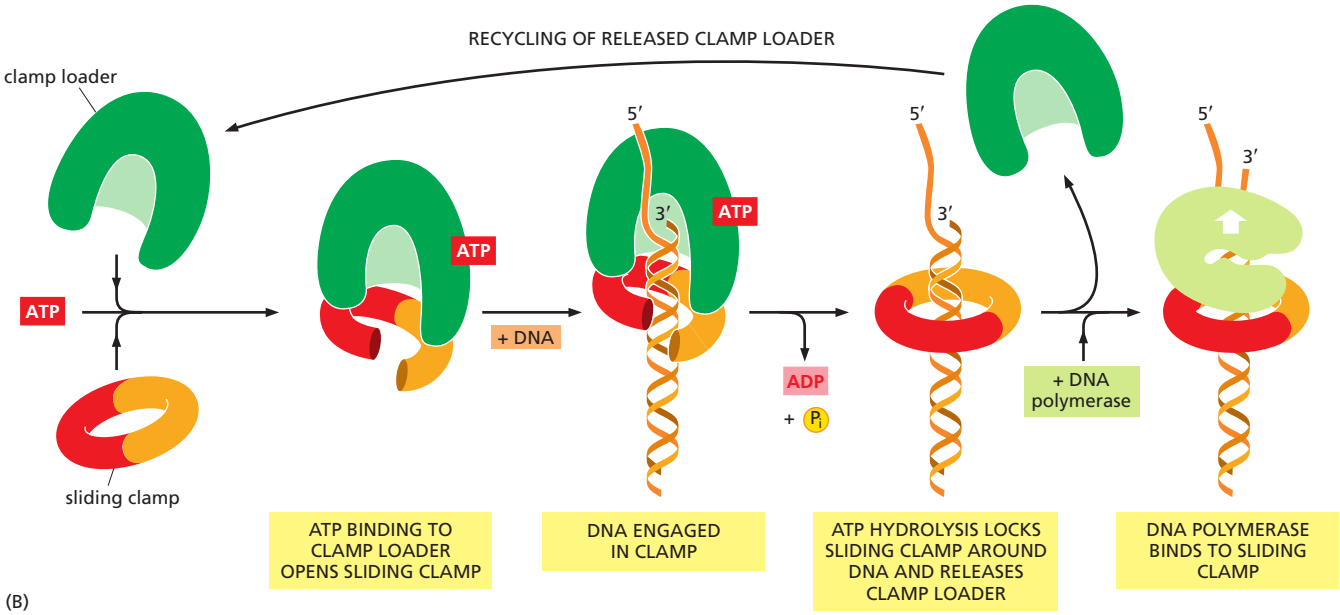


Figure 5–16 Human single-strand binding protein bound to DNA. (A) Front view of the two DNA-binding domains of the protein (called RPA) which cover a total of eight nucleotides. Note that the DNA bases remain exposed in this protein–DNA complex. (B) Diagram showing the three-dimensional structure, with the DNA strand *(orange)* viewed end-on. (PDB code: 1JMC.)

Figure 5–17 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 (**Movie 5.3**). (B) Schematic illustration showing how the clamp (with *red* and *yellow* subunits) is loaded onto DNA to serve as a tether for a moving DNA polymerase molecule. The structure of the clamp loader *(dark green)* resembles a screw nut, with its threads matching the grooves of double-stranded DNA. The loader binds to a free clamp molecule, forcing a gap in its ring of subunits so that this ring is able to slip around DNA. The clamp loader, thanks to its screw-nut structure, recognises the region of DNA that is double-stranded and latches onto it, tightening around the complex of a template strand with a freshly synthesized elongating (primer) strand. It carries the clamp along this double-stranded region until it encounters the 3′ end of the primer, at which point the loader hydrolyzes ATP and releases the clamp, allowing it to close around the DNA and bind to DNA polymerase. 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 polymerase so that, on the lagging strand, it is ready to assemble a new clamp at the start of each new Okazaki fragment (see Figure 5–18). (A, from X.P. Kong et al., *Cell* 69:425–437, 1992. With permission from Elsevier; B, adapted from B.A. Kelch et al., *Science* 334:1675–1680, 2011. With permission from AAAS. PDB code: 3BEP.)





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 DNA. This complex can be likened to a tiny sewing machine composed of protein parts and powered by nucleoside triphosphate hydrolysis. 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 eukaryotes, as we see below.

Figure 5–18 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 5–18A. 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

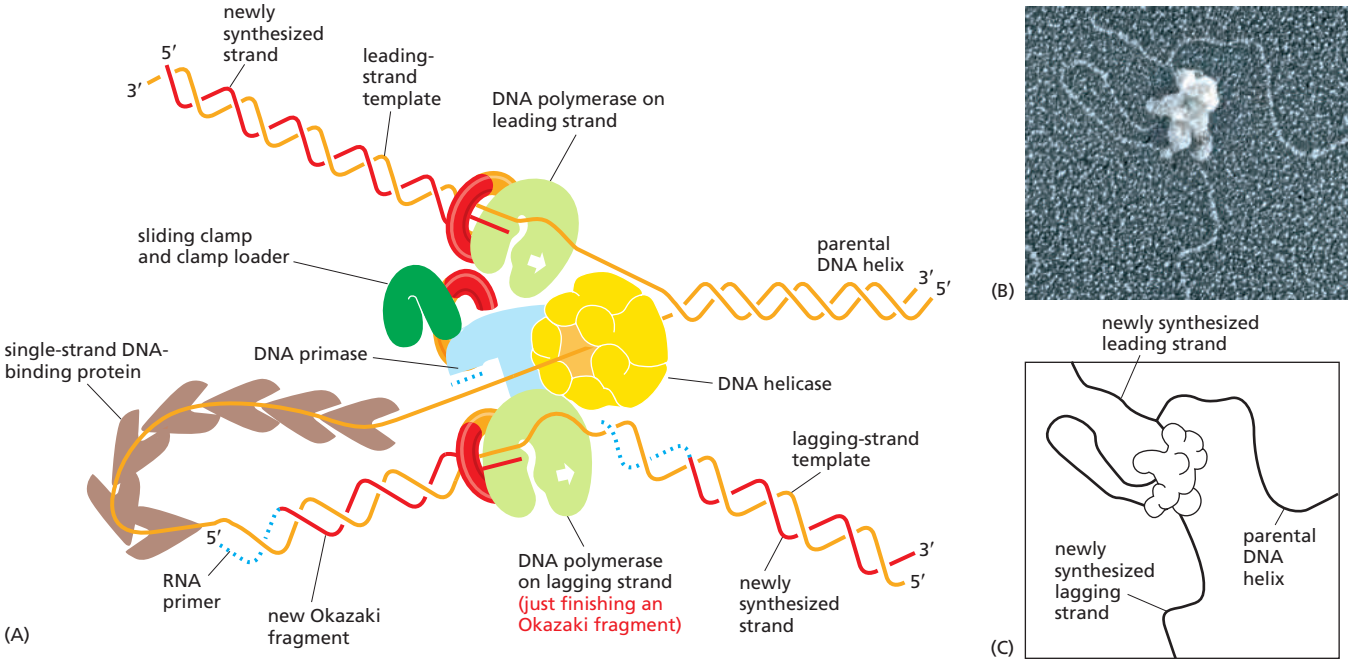


Figure 5–18 A bacterial replication fork. (A) This schematic diagram shows a current view of the arrangement of replication proteins at a replication fork when DNA is being synthesized. The lagging-strand DNA is folded to bring the lagging-strand DNA polymerase molecule into a complex with the leading-strand 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 is just being synthesized. 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 (**Movie 5.4** and **Movie 5.5**). (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.)

unit (total molecular mass >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. As discussed earlier, this RNA is removed and the resulting gap is filled in by DNA repair enzymes that operate behind the replication fork (see Figure 5–11).

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 30 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 consists of those with 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 newly synthesized strand, excision of the portion containing the mismatch, 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 to 1000 (see Table 5–1, p. 244).

A similar mismatch proofreading system functions in eukaryotic cells but uses a different strategy to distinguish the new strand from the old (**Figure 5–19**). Newly synthesized lagging-strand DNA transiently contains *nicks* (before they are sealed by DNA ligase) and such nicks (also called *single-strand breaks*) provide the signal that directs the mismatch proofreading system to the appropriate strand. This strategy also requires that the newly synthesized DNA on the leading strand be transiently nicked; how this occurs is uncertain.

The importance of mismatch proofreading 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 one 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 (see pp. 1096–1097),

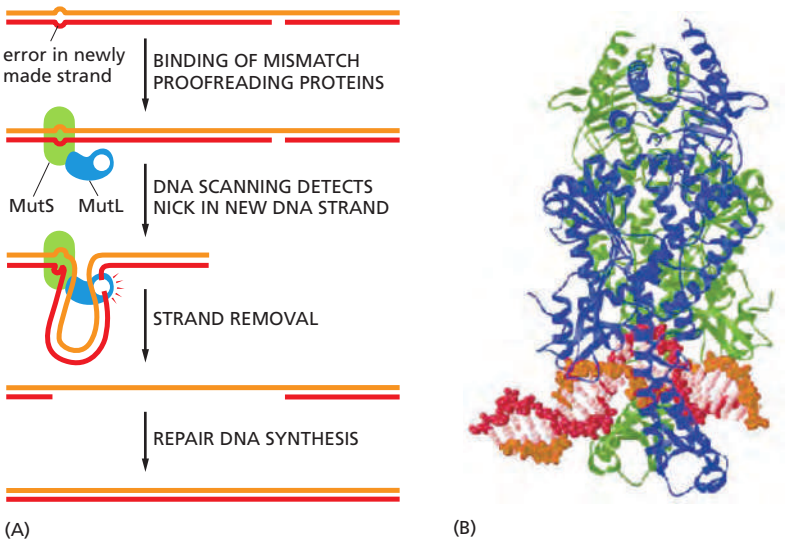


Figure 5–19 Strand-directed mismatch repair. (A) The two proteins shown are present in both bacteria and eukaryotic cells: MutS binds specifically to a mismatched base pair, while MutL scans the nearby DNA for a nick. Once MutL finds a nick, it triggers the degradation of the nicked strand all the way back through the mismatch. Because nicks are largely confined to newly replicated strands in eukaryotes, replication errors are selectively removed. In bacteria, an additional protein in the complex (MutH) nicks unmethylated (and therefore newly replicated) GATC sequences, thereby beginning the process illustrated here. In eukaryotes, MutL contains a DNA nicking activity that aids in the removal of the damaged strand. (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 with an abnormal base pair. (PDB code: 1EWQ.)

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.

DNA Topoisomerases Prevent DNA Tangling During Replication

As a replication fork moves along double-strand DNA, it creates what has been called the "winding problem." The two parental strands, which are wound around each other, must be unwound and separated for replication to occur. For every 10 nucleotide pairs replicated at the fork, one complete turn of the parental double helix must be unwound. In principle, this unwinding can be achieved by rapidly rotating the entire chromosome ahead of a moving fork; however, this is energetically highly unfavorable (particularly for long chromosomes) and, instead, the DNA in front of a replication fork becomes overwound (**Figure 5–20**). The overwinding, in turn, is continually relieved 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; 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 5–21**). 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

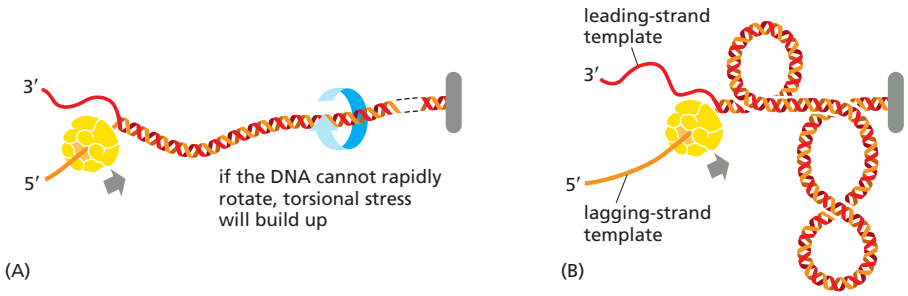


Figure 5–20 The "winding problem" that arises during DNA replication. (A) For a bacterial replication fork moving at 500 nucleotides per second, the parental DNA helix ahead of the fork must rotate at 50 revolutions per second. (B) If the ends of the DNA double helix remain fixed (or difficult to rotate), tension builds up in front of the replication fork as it becomes overwound. Some of this tension can be taken up by supercoiling, whereby the DNA double helix twists around itself (see Figure 6–19). However, if the tension continues to build up, the replication fork will eventually stop because further unwinding requires more energy than the helicase can provide. Note that in (A), the dotted line represents about 20 turns of DNA.

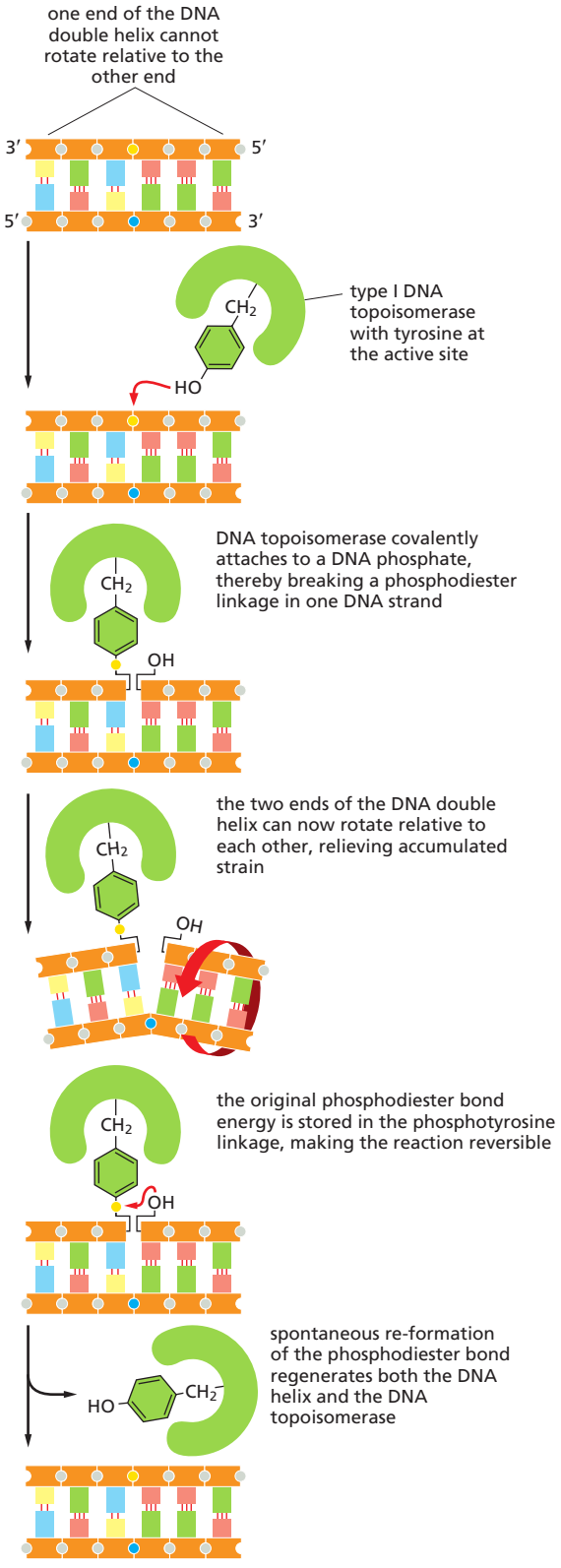


Figure 5–21 The reversible DNA nicking reaction catalyzed by a eukaryotic DNA topoisomerase I enzyme. As indicated, these enzymes transiently form a single covalent bond with DNA; this allows free rotation of the DNA around the covalent backbone bonds linked to the *blue* phosphate.

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 5–12).

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

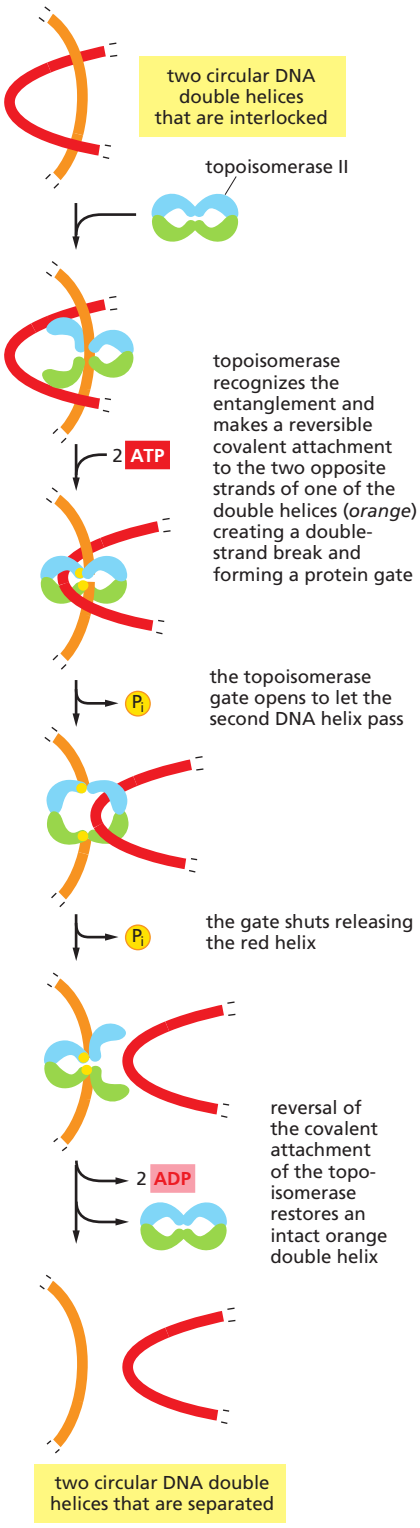


Figure 5–22 The DNA-helix-passing reaction catalyzed by DNA topoisomerase II. Unlike type I topoisomerases, type II enzymes hydrolyze ATP *(red),* which is needed to release and reset the enzyme after each cycle. Type II topoisomerases are largely confined to proliferating cells in eukaryotes; partly for that reason, they have been effective targets for anticancer drugs. Some of these drugs inhibit topoisomerase II at the third step in the figure and thereby produce high levels of double-strand breaks that kill rapidly dividing cells. The small *yellow circles* represent the phosphates in the DNA backbone that become covalently bonded to the topoisomerase (see Figure 5–21).

double-strand break in the helix. These enzymes are activated by sites on chromosomes where two double helices cross over each other such as those generated by supercoiling in front of a replication fork (see Figure 5–20). 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 opening; and (3) it then reseals the break and dissociates from the DNA. At crossover points generated by supercoiling, passage of the double helix through the gate occurs in the direction that will reduce supercoiling. In this way, type II topoisomerases can relieve the overwinding tension generated in front of a replication fork. Their reaction mechanism also allows type II DNA topoisomerases to efficiently separate two interlocked DNA circles (**Figure 5–22**).

Topoisomerase II also prevents the severe DNA tangling problems that would otherwise arise during DNA replication. This role is nicely illustrated by 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 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 Eukaryotes 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 eukaryotes, 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 eukaryotes.

There are more protein components in eukaryotic replication machines than there are in the bacterial analogs, even though the basic functions are the same. Thus, for example, the eukaryotic single-strand binding (SSB) protein is formed from three subunits, whereas only a single subunit is found in bacteria. Similarly, the eukaryotic DNA primase is incorporated into a multisubunit enzyme that also contains a 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 eukaryotic replicative DNA polymerases, Polδ and Polε, come into play: Polδ completes each Okazaki fragment on the lagging strand and Polε extends the leading strand. The increased complexity of eukaryotic replication machinery probably reflects

more elaborate controls. For example, the orderly maintenance of different cell types and tissues in animals and plants requires that DNA replication be tightly regulated. Moreover, eukaryotic DNA replication must be coordinated with the elaborate process of mitosis, as we discuss in Chapter 17.

As we see in the next section, the eukaryotic 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, as we will see, explains why new Okazaki fragments are synthesized on the lagging strand at intervals of 100–200 nucleotides in eukaryotes, 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 eukaryotic replication forks move only about one-tenth as fast as bacterial replication forks.

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 an intact, double-strand 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 eukaryotic cells must overcome. These include the need to replicate the enormously long DNA molecules found in eukaryotic 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 begin DNA replication, the double helix must first 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-strand DNA and pry the two strands apart, breaking the hydrogen bonds between the bases.

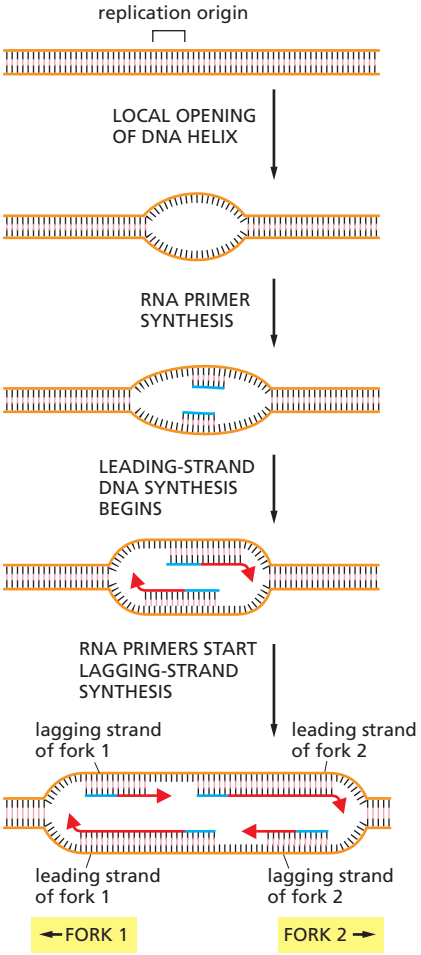


Figure 5–23 A replication bubble formed by replication-fork initiation. This diagram outlines the major steps in the initiation of replication forks at replication origins. The structure formed at the last step, in which both strands of the parental DNA helix have been separated from each other and serve as templates for DNA synthesis, is called a *replication bubble.*

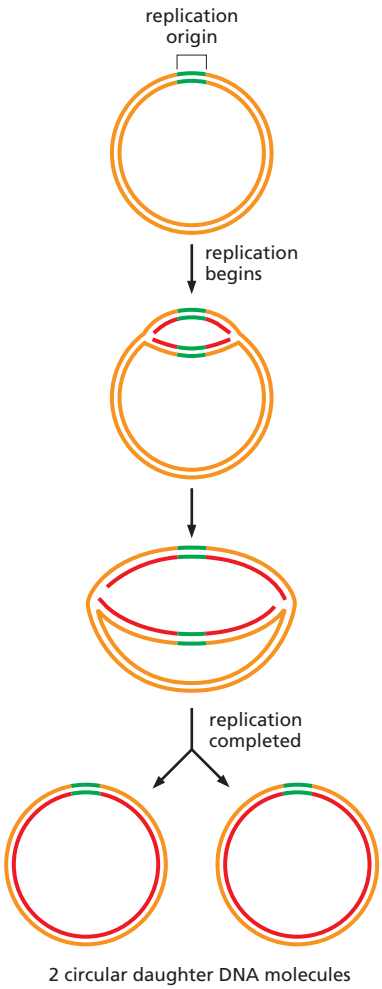


Figure 5–24 DNA replication of a bacterial genome. It takes *E. coli* about 30 minutes to duplicate its genome of 4.6×10^6 nucleotide pairs. For simplicity, no Okazaki fragments are shown on the lagging strand. What happens as the two replication forks approach each other and collide at the end of the replication cycle is not well understood, although the replication machines are disassembled as part of the process.

The positions at which the DNA helix is first opened are called **replication origins** (**Figure 5–23**). 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 base pairs are typically found at replication origins.

Although the basic process of replication-fork initiation depicted in Figure 5–23 is fundamentally the same for bacteria and eukaryotes, 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 eukaryotes.

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 1000 nucleotides per second) in opposite directions until they meet up roughly halfway around the chromosome (**Figure 5–24**). 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 (in their ATP-bound state) bind in multiple copies to specific DNA sites located at the replication origin, wrapping the DNA around the proteins to form a large protein–DNA complex that destabilizes the adjacent double helix. This complex then attracts two DNA helicases, each bound to a helicase loader, and these are placed around adjacent DNA single strands 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 helicases are loaded, the loaders dissociate and the helicases begin to unwind DNA, exposing enough single-strand DNA for DNA primase to synthesize the first RNA primers (**Figure 5–25**). This quickly leads to the assembly of remaining proteins to create two replication forks, with replication machines that move, with respect to the replication origin, in opposite directions. They 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. Initiation is also controlled to ensure that only one round of DNA replication occurs for each cell division. After replication is initiated, the initiator protein is inactivated by hydrolysis of its bound ATP molecule, and the origin of replication experiences a "refractory period." The refractory period is caused by a delay in the methylation of newly incorporated A nucleotides in the origin (**Figure 5–26**). Initiation cannot occur again until the A's are methylated and the initiator protein is restored to its ATP-bound state.

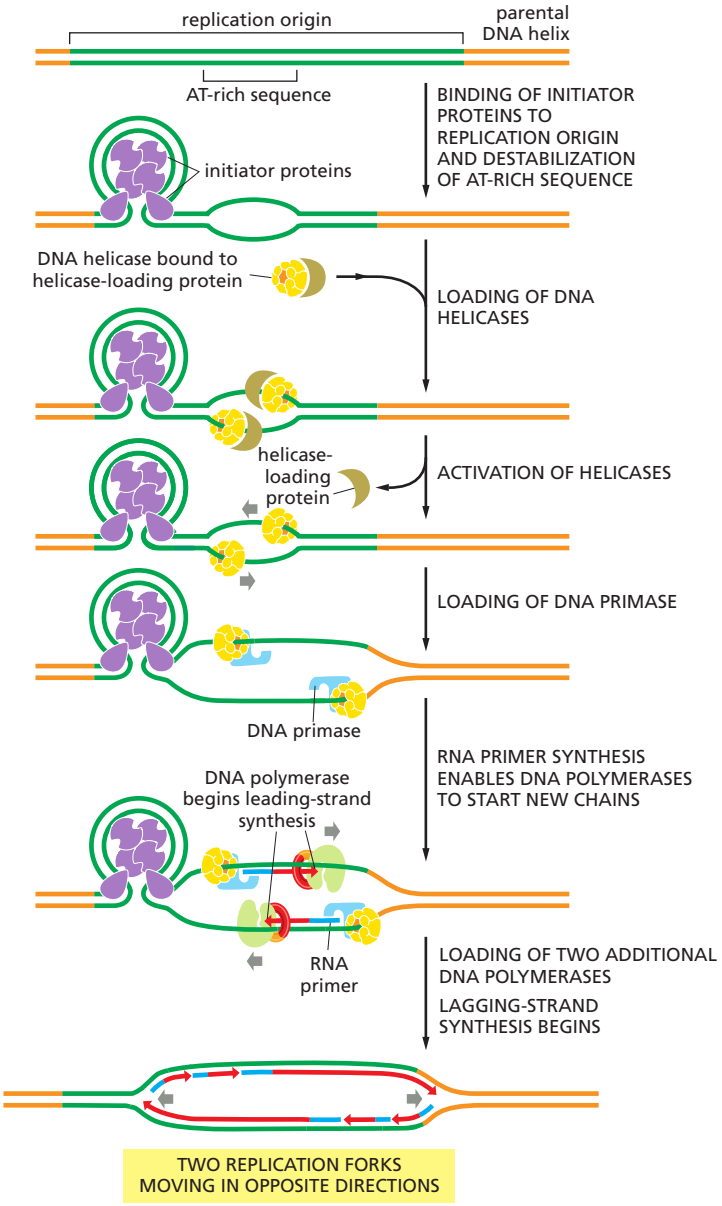


Figure 5–25 The proteins that initiate DNA replication in bacteria. The mechanism shown was established by studies *in vitro* with mixtures of highly purified proteins. For *E. coli* DNA replication, the major initiator protein, the helicase, and the primase are the dnaA, dnaB, and dnaG proteins, respectively. In the first step, several molecules of the initiator protein bind to specific DNA sequences at the replication origin and destabilize the double helix by forming a compact structure in which the DNA is tightly wrapped around the protein. Next, two helicases are brought in by helicase-loading proteins (the dnaC proteins), which inhibit the helicases until they are properly loaded at the replication origin. Helicase-loading proteins prevent the replicative DNA helices from inappropriately entering other single-strand stretches of DNA in the bacterial genome. Aided by single-strand binding protein (not shown), the loaded helicases open up the DNA, thereby enabling primases to enter and synthesize initial primers. In subsequent steps, two complete replication forks are assembled at the origin and move off in opposite directions. The initiator proteins are displaced as the left-hand fork moves through them (not shown).

Eukaryotic 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 30 minutes. Because of the much greater size of most eukaryotic chromosomes, a different strategy is required to allow their replication in a timely manner.

A method for determining the general pattern of eukaryotic 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 stretched on the surface of a glass slide coated with a photographic emulsion. Development of the emulsion reveals the pattern of labeled DNA through 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.

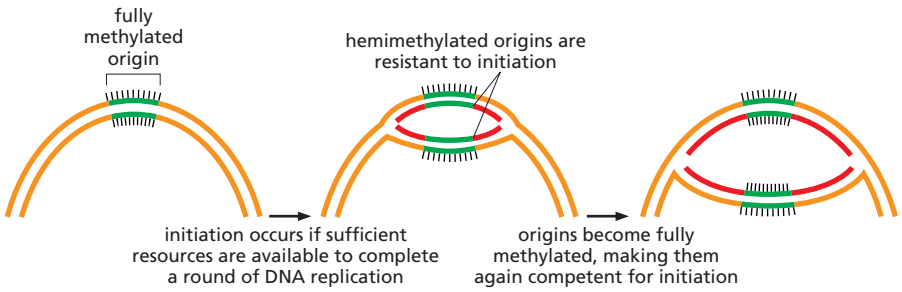


Figure 5–26 Methylation of the *E. coli* replication origin creates a refractory period for DNA initiation. DNA methylation occurs at GATC sequences, 11 of which are found in the origin of replication (spanning approximately 250 nucleotide pairs). In its hemimethylated state, the origin of replication is bound by an inhibitor protein (Seq A, not shown), which blocks the ability of the initiator proteins to unwind the origin DNA. Eventually (about 15 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.

In this way, both the rate and the direction of replication-fork movement can be determined (**Figure 5–27**). From the rate at which tracks of replicated DNA increase in length with increasing labeling time, the eukaryotic replication forks are estimated to travel at about 50 nucleotides per second. This is approximately twentyfold slower than 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 35 days) 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, belonging to separate replication bubbles, are moving simultaneously on each eukaryotic chromosome.

Much faster and more sophisticated methods now exist for monitoring DNA replication initiation and tracking the movement of DNA replication forks across whole genomes. One approach uses DNA microarrays—grids the size of a postage stamp studded with hundreds 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 replicating cells 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 across an entire genome (**Figure 5–28**).

Experiments of this type have shown the following: (1) Approximately 30,000–50,000 origins of replication are used each time a human cell divides. (2) The human genome has many more (perhaps tenfold more) potential origins than this, and different cell types use different sets of origins. This may allow a cell to coordinate its active origins with other features of its chromosomes such as which

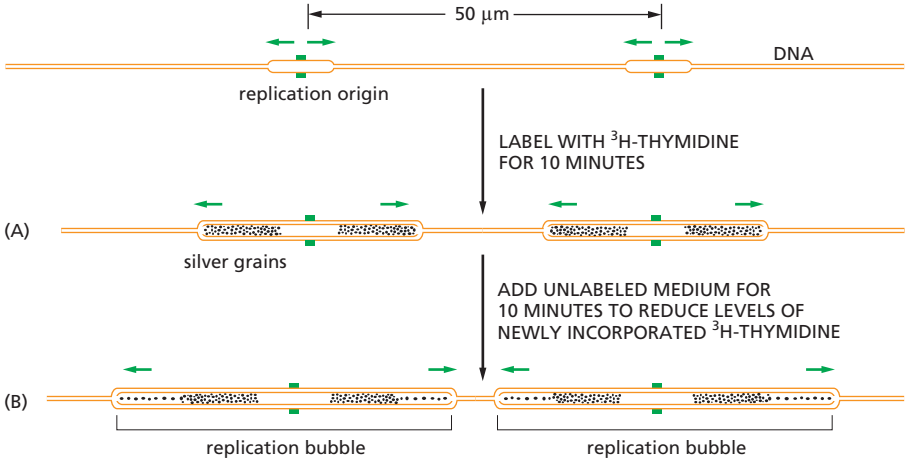


Figure 5–27 The experiments that demonstrated the pattern in which replication forks are formed and move on eukaryotic chromosomes. The new DNA made in human cells in culture was labeled briefly with a pulse of highly radioactive thymidine (^{3}H-thymidine). (A) In this experiment, the cells were lysed, 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 5–23). 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.

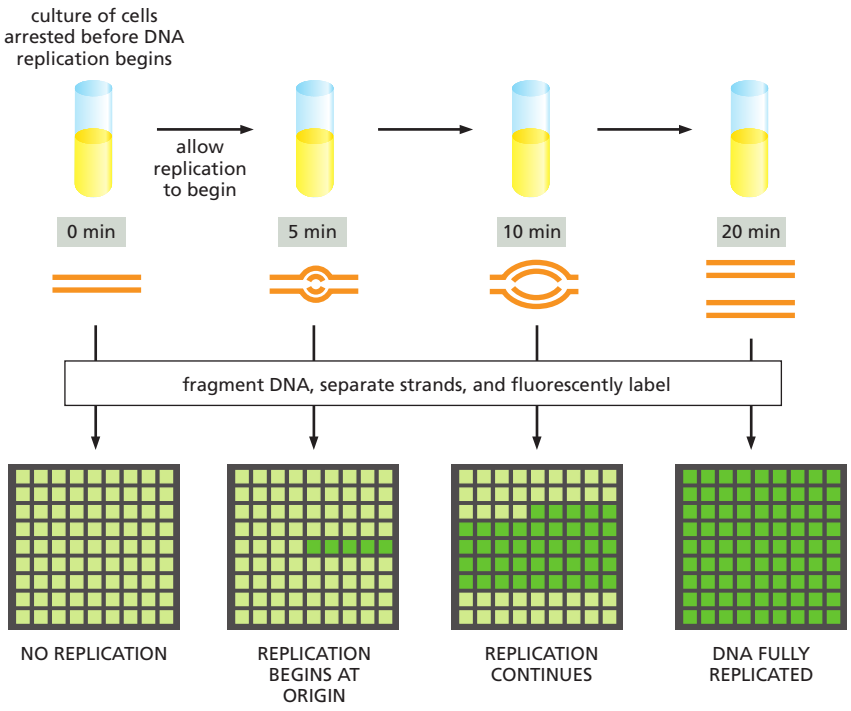


Figure 5–28 Use of DNA microarrays to monitor the formation and progress of replication forks. For this experiment, a population of cells is synchronized so that they all begin replication at the same time. DNA is collected and hybridized to the microarray; DNA that has been replicated once gives a hybridization 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 chromosome arranged left to right, top to bottom. Only 81 spots are shown here, but the actual arrays contain hundreds of thousands of sequences that span an entire genome. As can be seen, replication begins at an origin and proceeds bidirectionally. For simplicity, only one origin is shown here. In human cells, replication begins at 30,000–50,000 origins located throughout the genome. Using this approach it is possible to observe the formation and progress of every replication fork across a genome.

genes are being expressed. The excess origins also provide "backups" in case a primary origin fails. (3) 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 Eukaryotes, DNA Replication Takes Place During Only One Part of the Cell Cycle

When growing rapidly, bacteria replicate their DNA nearly continuously. In contrast, DNA replication in most eukaryotic cells occurs only during a specific part of the cell-division cycle, called the *DNA synthesis phase* or **S phase** (Figure 5–29). In a mammalian cell, the S phase typically lasts for about 8 hours; in simpler eukaryotic 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 we 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.

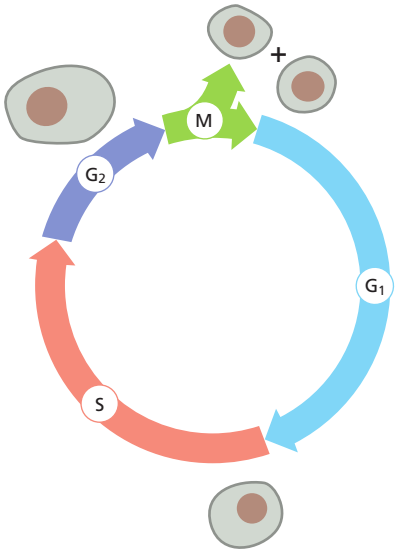


Figure 5–29 The four successive phases of a standard eukaryotic cell cycle. During the G_1, S, and G_2 phases, the cell grows continuously. During M phase growth stops, the nucleus divides, and the cell divides in two. DNA replication is confined to the part of the cell cycle known as S phase. G_1 is 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; indeed, replication origins are activated in clusters of about 50 adjacent replication origins, each of which is replicated during only a small part of the total S-phase interval.

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, where most transcription occurs, has a less condensed conformation. 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.

Once initiated, however, 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.

A Large Multisubunit Complex Binds to Eukaryotic Origins of Replication

Having seen that a eukaryotic 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. We shall see that this is the case for the single-cell budding yeast *S. cerevisiae*, but it appears not to be strictly true for most other eukaryotes.

For budding yeast, the location of every origin of replication on each chromosome has been determined. The particular chromosome shown in **Figure 5–30**—chromosome III from *S. cerevisiae*—is one of the smallest 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, but only a subset of these origins is used by a given cell. Nonetheless, this chromosome can be replicated in about 15 minutes.

The minimal DNA sequence required for directing DNA replication initiation in *S. cerevisiae* has been determined by taking a segment of DNA that spans an origin of replication and testing smaller and smaller DNA fragments for their ability to function as origins. Most DNA sequences that can serve as an origin of replication are 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 melt; and (3) at least one binding site for proteins that facilitate ORC binding, probably by adjusting chromatin structure.

In bacteria, once the initiator protein is properly bound to the single origin of replication, the assembly of the replication forks seems to follow more or less automatically. In eukaryotes, the situation is significantly different because of a profound problem eukaryotes have in replicating chromosomes: 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 sequential manner in which the replicative helicase is first loaded onto origins and is then activated to initiate DNA replication. This matter is discussed in detail in Chapter 17, where we consider the machinery that underlies the cell-division cycle. In brief, during G_1 phase, the replicative helicases are loaded onto DNA next to ORC to create a *prereplicative complex*. Then, upon passage from G_1 phase to S phase, specialized protein kinases come into play to activate the helicases. The resulting opening of the double helix allows the loading of the remaining replication proteins, including the DNA polymerases.

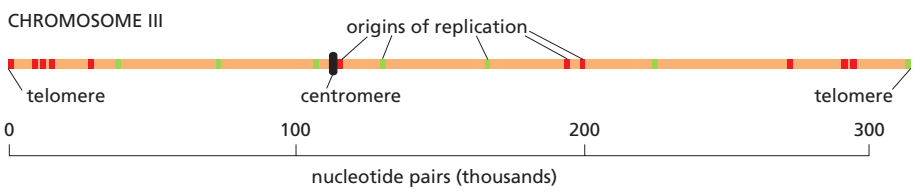


Figure 5–30 The origins of DNA replication on chromosome III of the yeast *S. cerevisiae*. This chromosome, one of the smallest eukaryotic chromosomes known, carries a total of 180 genes. As indicated, it contains 18 replication origins, although they are used with different frequencies. Those in *red* are typically used in less than 10% of cell divisions, while those in *green* are used about 90% of the time.

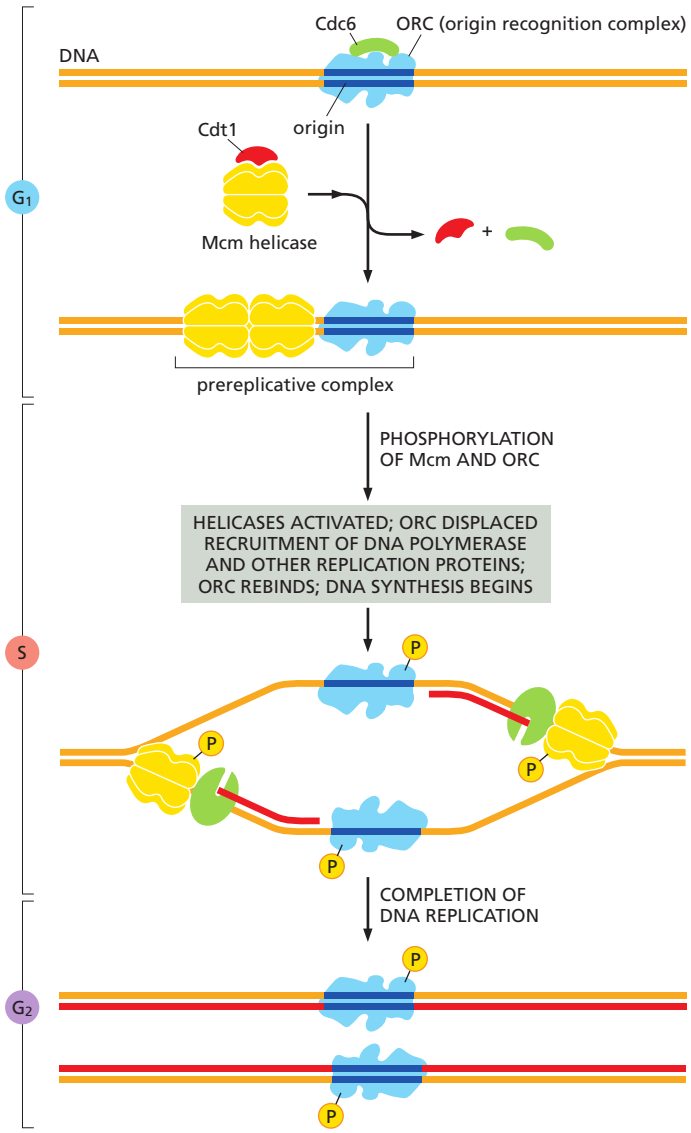


Figure 5–31 DNA replication initiation in eukaryotes. This mechanism ensures that each origin of replication is activated only once per cell cycle. An origin of replication can be used only if a prereplicative complex forms there in G_1 phase. At the beginning of S phase, specialized kinases phosphorylate Mcm and ORC, activating the former and inactivating the latter. A new prereplicative complex cannot form at the origin until the cell progresses to the next G_1 phase, when the bound ORC has been dephosphorylated. Note that the eukaryotic Mcm helicase moves along the leading-strand template, whereas the bacterial helicase moves along the lagging-strand template (see Figure 5–25). As the forks begin to move, ORC is displaced, and new ORCs rapidly bind to the newly replicated origins.

The protein kinases that trigger DNA replication simultaneously prevent assembly of new prereplicative complexes until the next M phase resets the entire cycle (for details, see pp. 974–975). They do this, in part, by phosphorylating ORC, rendering it unable to accept new helicases. This strategy provides a single window of opportunity for prereplicative complexes to form (G_1 phase, when kinase activity is low) and a second window for them to be activated and subsequently disassembled (S phase, when kinase 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 during each cell cycle.

Features of the Human Genome That Specify Origins of Replication Remain to Be Discovered

Compared with the situation in budding yeast, the determinants of replication origins in other eukaryotes have been difficult to discover. 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. However, comparisons of such DNA sequences have not revealed specific DNA sequences that mark origins of replication.

Despite this, a human ORC that is very similar to the yeast ORC binds to origins of replication and initiates DNA replication in humans. 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 are similar in outline, but chromatin structure, transcriptional activity, or some property of the genome other than a specific DNA sequence has the central role in attracting ORC and specifying mammalian origins of replication. These ideas could also help to explain how a given mammalian cell chooses which of the many possible origins to use when it replicates its genome and how this choice could differ from cell to cell. Clearly, we have a great deal to discover about the fundamental process of DNA replication initiation.

New Nucleosomes Are Assembled Behind the Replication Fork

Several additional aspects of DNA replication are specific to eukaryotes. As discussed in Chapter 4, eukaryotic 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 eukaryotic 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, histones are synthesized mainly in S phase, when the level of histone mRNA increases about fiftyfold as a result of both increased transcription and 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 pass through the parental nucleosomes. In the cell, efficient replication requires chromatin remodeling complexes (discussed in Chapter 4) to destabilize the DNA–histone interfaces. Aided by such complexes, replication forks can transit even highly condensed chromatin efficiently.

As a replication fork passes through chromatin, the histones are transiently displaced leaving about 600 nucleotide pairs of non-nucleosomal DNA in its wake. The reestablishment of nucleosomes behind a moving fork occurs in an intriguing way. When a nucleosome is traversed by a replication fork, the histone octamer appears to be broken into an H3-H4 tetramer and two H2A-H2B dimers (discussed in Chapter 4). The H3-H4 tetramer remains loosely associated with DNA and is distributed at random to one or the other daughter duplex, but the H2A-H2B dimers are released completely from DNA. Freshly made H3-H4 tetramers are added to the newly synthesized DNA to fill in the "spaces," and H2A-H2B dimers—half of which are old and half new—are then added at random to complete the nucleosomes (**Figure 5–32**). The formation of new nucleosomes behind a replication fork has an important consequence for the process of DNA replication itself. As DNA polymerase δ discontinuously synthesizes the lagging strand (see pp. 253–254), the length of each Okazaki fragment is determined by the point at which DNA polymerase δ is blocked by a newly formed nucleosome. This tight coupling between nucleosome duplication and DNA replication explains why the length of Okazaki fragments in eukaryotes (~200 nucleotides) is approximately the same as the nucleosome repeat length.

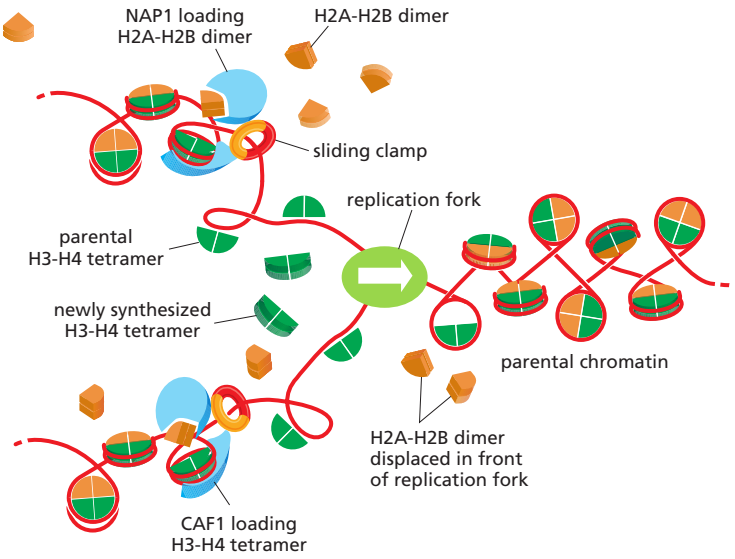


Figure 5–32 Formation of nucleosomes behind a replication fork. Parental H3-H4 tetramers are distributed at random to the daughter DNA molecules, with roughly equal numbers inherited by each daughter. In contrast, H2A-H2B dimers are released from the DNA as the replication fork passes. This release begins just in front of the replication fork and is facilitated by chromatin remodeling complexes that move with the fork. Histone chaperones (NAP1 and CAF1) restore the full complement of histones to daughter molecules using both parental and newly synthesized histones. Although some daughter nucleosomes contain only parental histones or only newly synthesized histones, most are hybrids of old and new. For simplicity, the DNA double helix shown as a single *red* line. (Adapted from J.D. Watson et al., Molecular Biology of the Gene, 5th ed. Cold Spring Harbor: Cold Spring Harbor Laboratory Press, 2004.)

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. The histone chaperones, along with their cargoes, are directed to newly replicated DNA through a specific interaction with the eukaryotic sliding clamp called *PCNA* (see Figure 5–32B). These clamps are left behind moving replication forks and remain on the DNA long enough for the histone chaperones to complete their tasks.

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. The final RNA primer synthesized on the lagging-strand template cannot be replaced by DNA because there is no 3′-OH end available for the repair polymerase. Without a mechanism to deal with this problem, DNA would be lost from the ends of all chromosomes each time a cell divides.

Bacteria solve this "end-replication" problem by having circular DNA molecules as chromosomes (see Figure 5–24). Eukaryotes solve it in a different way: they have specialized nucleotide sequences at the ends of their chromosomes that are incorporated into structures called *telomeres* (discussed in 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 5–33**). The enzymatic portion of telomerase resembles other *reverse transcriptases*, proteins that synthesize DNA using an RNA template, although, in this case, the telomerase RNA also contributes functional groups to make the catalysis more efficient. 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 5–34**).

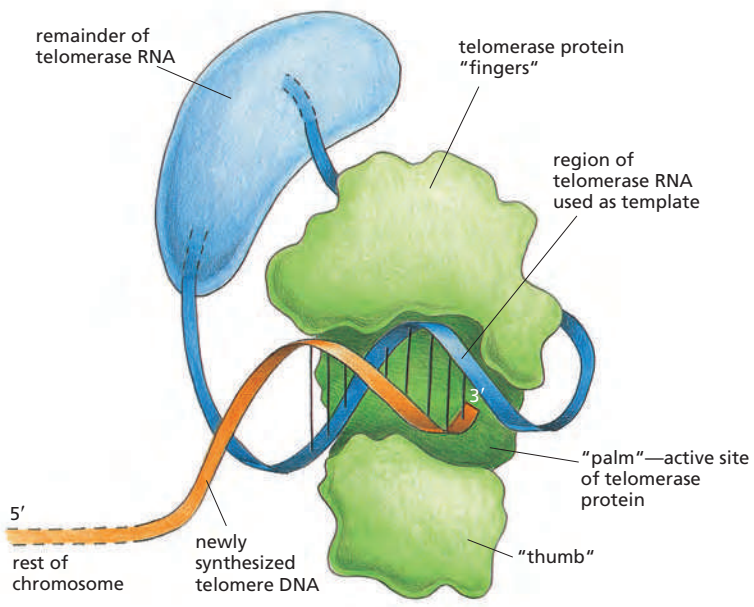


Figure 5–33 Structure of a portion of telomerase. Telomerase is a large protein–RNA complex. The RNA *(blue)* contains a templating sequence for synthesizing new DNA telomere repeats. The synthesis reaction itself is carried out by the reverse transcriptase domain of the protein, shown in *green.* A 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. Telomerase also has several additional protein domains (not shown) that are needed to assemble the enzyme at the ends of chromosomes. (Modified from J. Lingner and T.R. Cech, *Curr. Opin. Genet. Dev.* 8:226–232, 1998. With permission from Elsevier.)

Telomeres Are Packaged Into Specialized Structures That Protect the Ends of Chromosomes

The ends of chromosomes present cells with an additional problem. As we will see in the next part of this chapter, when a chromosome is accidently broken, the break is rapidly repaired (see Figure 5–45). Telomeres must clearly be distinguished from these accidental breaks; otherwise the cell will attempt to "repair" telomeres, causing chromosome fusions and other genetic abnormalities. Telomeres have several features to prevent this from happening.

A specialized nuclease chews back the 5′ end of a telomere leaving a protruding single-strand end. This protruding end—in combination with the GGGTTA repeats in telomeres—attracts a group of proteins that form a protective chromosome cap known as *shelterin*. In particular, shelterin "hides" telomeres from the cell's damage detectors that continually monitor DNA. When human telomeres are artificially cross-linked and viewed by electron microscopy, structures known as "t-loops" are observed in which the protruding end of the telomere loops back and tucks itself into the duplex DNA of the telomere repeat sequence (**Figure 5–35**). It is believed that t-loops are regulated by shelterin and provide additional protection for the ends of chromosomes.

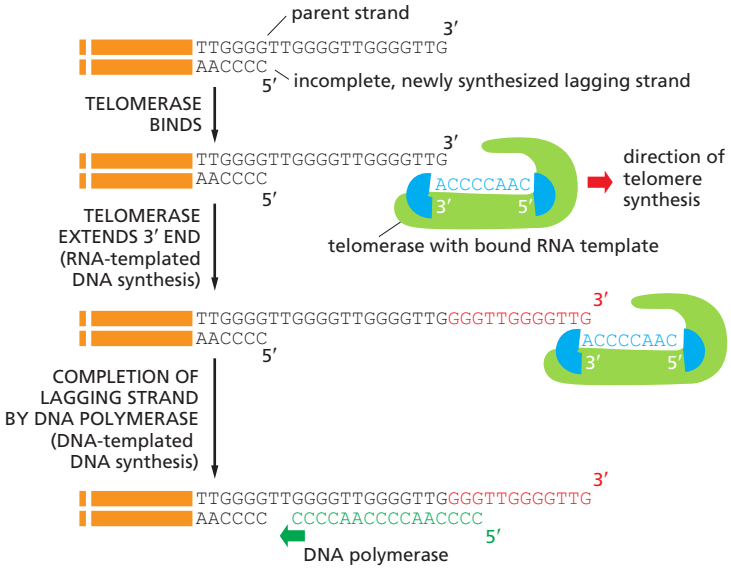


Figure 5–34 Telomere replication. Shown here are the reactions that synthesize the repeating sequences that form the ends of the chromosomes (telomeres) of diverse eukaryotic organisms. The 3′ end of the parental DNA strand is extended by RNA-templated DNA synthesis; this allows the incomplete daughter 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 (**Movie 5.6**). The telomere sequence illustrated is that of the ciliate *Tetrahymena,* in which these reactions were first discovered.

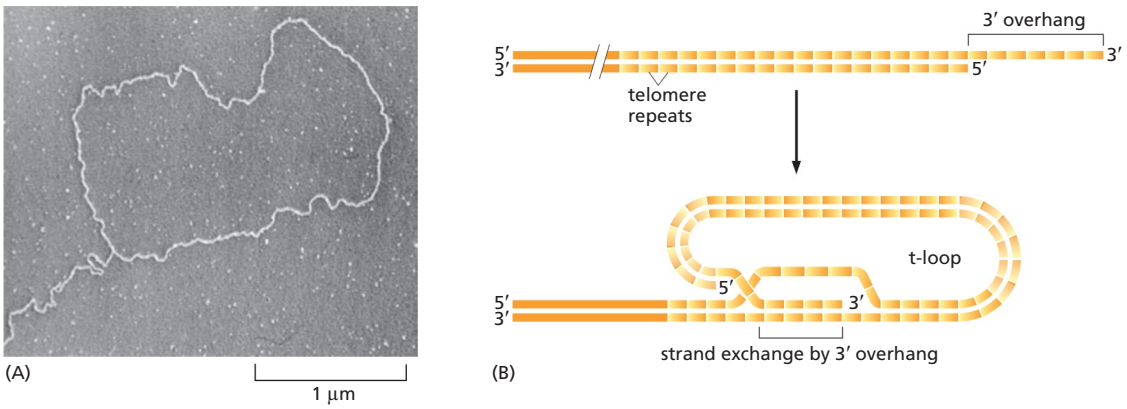


Figure 5–35 A t-loop at the end of a mammalian chromosome. (A) Electron micrograph of the DNA at the end of an interphase human chromosome. The chromosome was fixed, deproteinated, and artificially thickened before viewing. The loop seen here is approximately 15,000 nucleotide pairs in length. (B) Structure of a t-loop. The insertion of the single-strand 3′ end into the duplex repeats is carried out, and the structure maintained, by specialized proteins. (From J.D. Griffith et al., *Cell* 97:503–514, 1999. With permission from Elsevier.)

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, many cells have homeostatic mechanisms that maintain the number of these repeats within a limited range (**Figure 5–36**).

In most of the dividing somatic cells of humans, however, telomeres gradually shorten, and it has been proposed that this provides a counting mechanism that helps prevent the unlimited proliferation of wayward cells in adult tissues. In its simplest form, this idea holds that our somatic cells start off in the embryo with a full complement of telomeric repeats. These are then eroded to different extents in different cell types. Some stem cells, notably those in tissues that must be replenished at a high rate throughout life—bone marrow or gut lining, for example—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 descendant cells will inherit chromosomes that lack telomere function, and, as a result of this defect, activate a DNA-damage response causing them to 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.

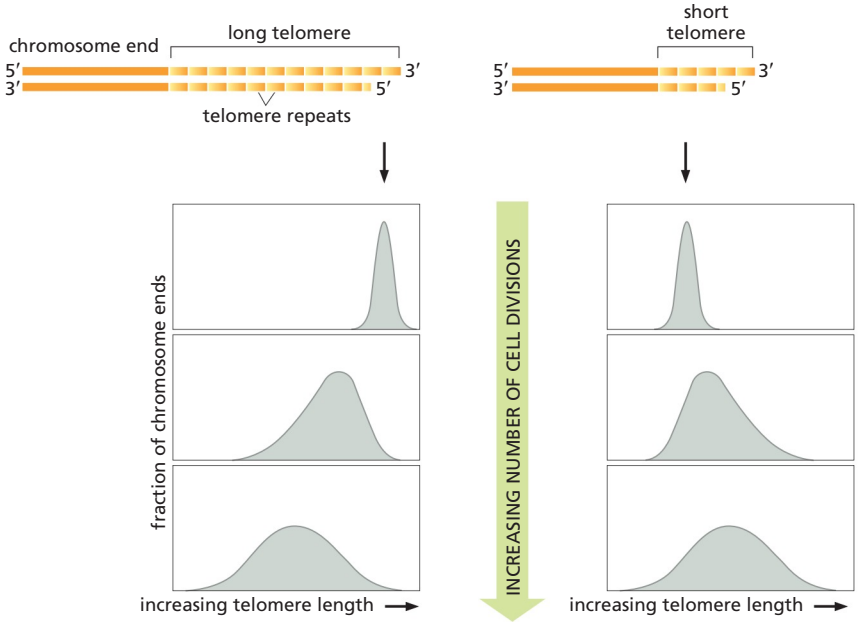


Figure 5–36 A demonstration that yeast cells control the length of their telomeres. In this experiment, the telomere at one end of a particular chromosome is artificially made either longer *(left)* or shorter *(right)* than average. After many cell divisions, the chromosome recovers, showing an average 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 germ-line cells of animals.

The idea that telomere length acts as a "measuring stick" to count cell divisions and thereby regulate the lifetime of the cell lineage 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 cell 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 has been proposed that this type of control on cell proliferation may contribute to 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 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. Clearly, the use of telomere shortening as a regulating mechanism is not foolproof and, like many mechanisms in the cell, seems to strike a balance between benefit and risk.

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 eukaryotes, replication origins are specified by specific DNA sequences that are only several hundred nucleotide pairs long. In other eukaryotes, 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. Eukaryotic DNA replication takes place in only one part of the cell cycle, the S phase. The replication fork in eukaryotes moves about 10 times more slowly than the bacterial replication fork, and the much longer eukaryotic 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 eukaryotic 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.

Eukaryotes 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. Telomeres have specialized structures that distinguish them from broken ends of chromosomes, ensuring that they are not mistakenly repaired.

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 DNA continually suffers. 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 tens of 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 (less than 0.02%) accumulate as permanent mutations in the DNA sequence. 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 the enzymes that carry it out: several percent of the coding capacity of most genomes 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**). Thus, we saw previously that defects in a human gene whose product normally functions to repair the mismatched base pairs resulting from DNA replication errors can lead to an inherited predisposition to cancers of the colon and some other organs, reflecting an increased mutation rate. In another human disease,

TABLE 5–2 Some Inherited Human 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
Cockayne syndrome	UV sensitivity; developmental abnormalities	Coupling of nucleotide excision repair to transcription
XP variant	UV sensitivity, skin cancer	Translesion synthesis by DNA polymerase ν
Ataxia telangiectasia (AT)	Leukemia, lymphoma, γ-ray sensitivity, genome instability	ATM protein, a protein kinase activated by double-strand breaks
BRCA1	Breast and ovarian cancer	Repair by homologous recombination
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 used in repair
Bloom syndrome	Cancer at several sites, stunted growth, genome instability	DNA helicase needed for recombination
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

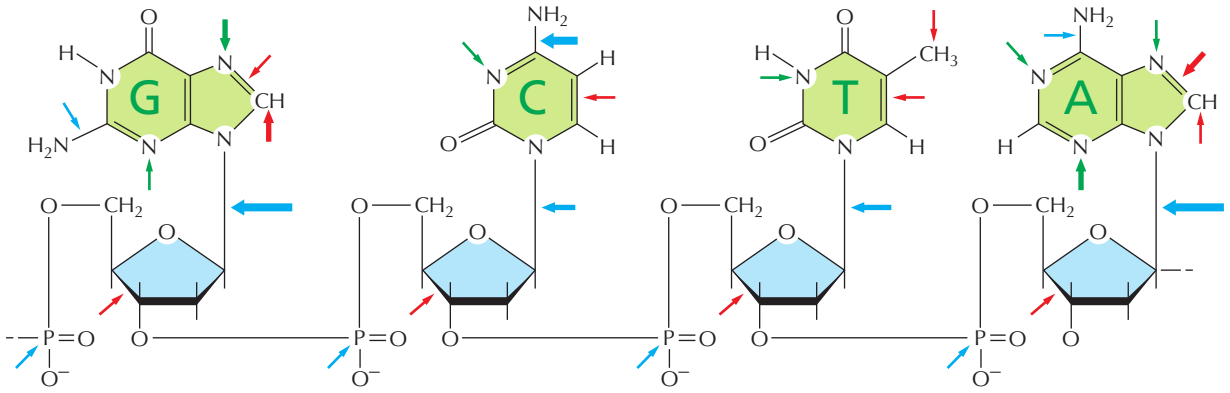


Figure 5–37 A summary of spontaneous alterations that require DNA repair. The sites on each nucleotide modified by spontaneous oxidative damage *(red arrows)*, hydrolytic attack *(blue arrows)*, and methylation *(green arrows)* are shown, with the width of each arrow indicating the relative frequency of each event (see Table 5–3). (After T. Lindahl, *Nature* 362:709–715, 1993. With permission from Macmillan Publishers Ltd.)

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 skin cancers. Finally, mutations in the *Brca1* and *Brca2* genes compromise a type of DNA repair known as *homologous recombination* and are a cause of hereditary breast and ovarian cancer.

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 5–37** and see **Table 5–3**). For example, the DNA of each

TABLE 5–3 Endogenous DNA Lesions Arising and Repaired in a Diploid Mammalian Cell in 24 Hours

DNA lesion	Number repaired in 24 h
Hydrolysis	
Depurination	18,000
Depyrimidination	600
Cytosine deamination	100
5-Methylcytosine deamination	10
Oxidation	
8-oxo G	1500
Ring-saturated pyrimidines (thymine glycol, cytosine hydrates)	2000
Lipid peroxidation products (M1G, etheno-A, etheno-C)	1000
Nonenzymatic methylation by *S*-adenosylmethionine	
7-Methylguanine	6000
3-Methyladenine	1200
Nonenzymatic methylation by nitrosated polyamines and peptides	
O^6-Methylguanine	20–100

The DNA lesions listed in the table are the result of the normal chemical reactions that take place in cells. Cells that are exposed to external chemicals and radiation suffer greater and more diverse forms of DNA damage. (From T. Lindahl and D.E. Barnes, *Cold Spring Harb. Symp. Quant. Biol.* 65:127–133, 2000.)

human cell loses about 18,000 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 5–38**). DNA bases are also occasionally damaged by an encounter with reactive metabolites produced in the cell, including reactive forms of oxygen and the high-energy methyl donor *S*-adenosylmethionine, 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 5–39**). 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 5–40**). 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.

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.

An indication of the importance of a double-strand helix to the safe storage of genetic information is that all cells use it; only a few small viruses use single-strand 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-strand 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.

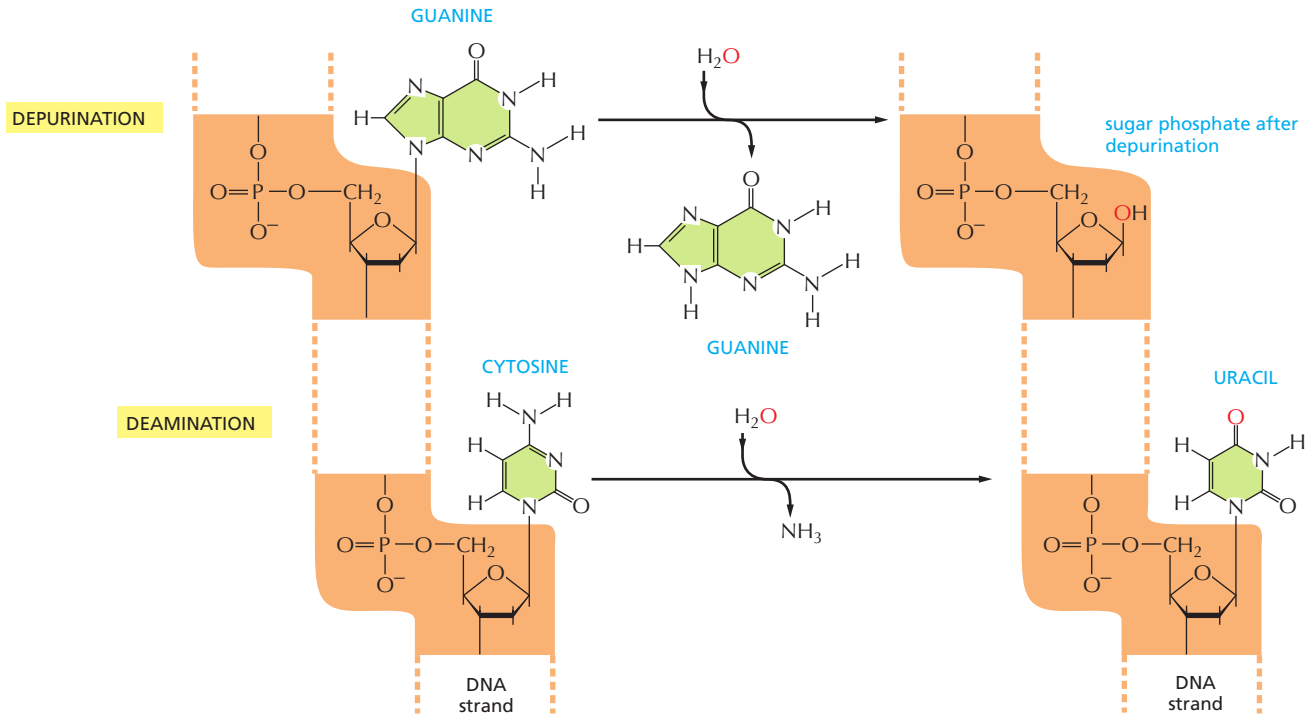


Figure 5–38 Depurination and deamination. These reactions are two of the most frequent spontaneous chemical reactions that create serious DNA damage in cells. Depurination can release guanine (shown here), as well as adenine, from DNA. The major type of deamination reaction converts cytosine to an altered 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.

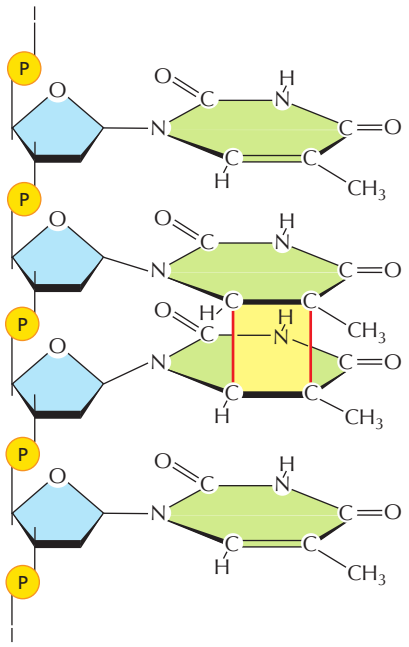


Figure 5–39 The most common type of thymine dimer. This type of damage occurs in the DNA of cells exposed to ultraviolet irradiation (as in sunlight). A similar dimer will form between any two neighboring pyrimidine bases (C or T residues) in DNA.

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 5–41** 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 5–12).

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" of the altered nucleotide from the helix, which allows the DNA glycosylase to probe all faces of the base for damage (**Figure 5–42**). 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 that 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 resulting gap is repaired (see Figure 5–41A). 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 5–41A.

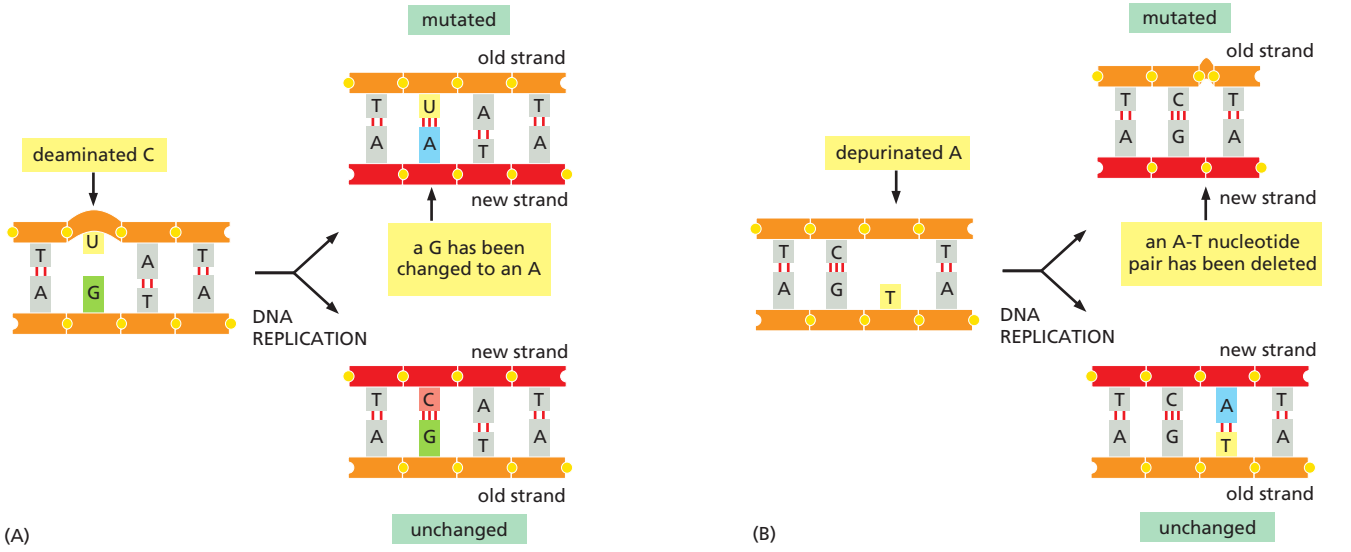


Figure 5–40 How chemical modifications of nucleotides produce mutations. (A) Deamination of cytosine, if uncorrected, results in the substitution of one base for another when the DNA is replicated. As shown in Figure 5–38, deamination of cytosine produces uracil. Uracil differs from cytosine in its base-pairing 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 5–37), if left uncorrected, also produce mutations when the DNA is replicated.

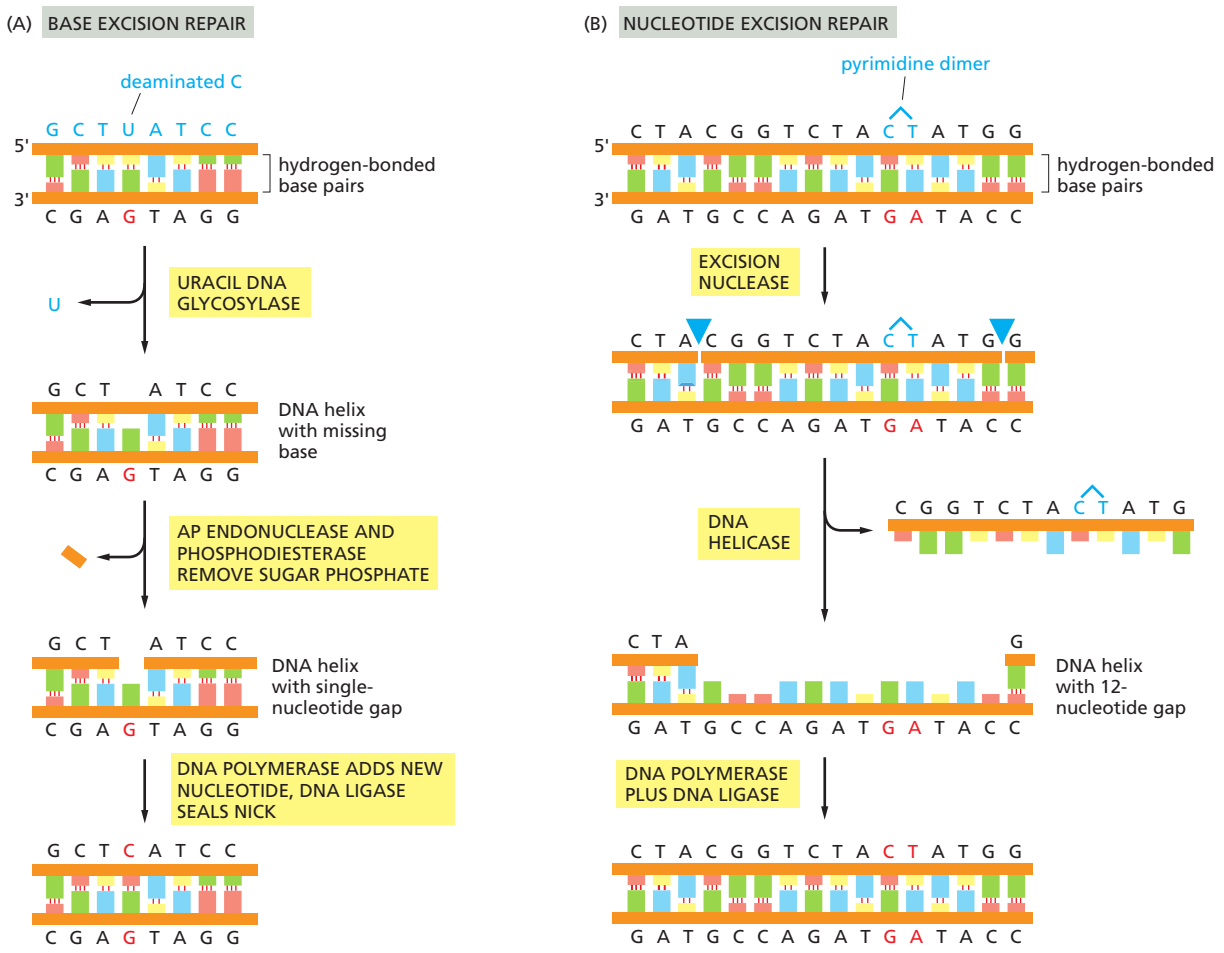


Figure 5– 41 A comparison of two major DNA repair pathways. (A) *Base excision repair*. This pathway starts with a DNA glycosylase. Here, the enzyme uracil DNA glycosylase removes an accidentally deaminated cytosine in DNA. After the action of this glycosylase (or another DNA glycosylase that recognizes 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 5–39), 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.

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, found in tobacco smoke, coal tar, and diesel exhaust), 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 lesion, it cleaves the phosphodiester backbone of the abnormal strand on both sides of the distortion, and a DNA helicase peels away the single-strand oligonucleotide containing the lesion. The large gap produced in the DNA helix is then repaired by DNA polymerase and DNA ligase (see Figure 5–41B).

An alternative to base and nucleotide excision repair processes is direct chemical reversal of DNA damage, and this strategy is selectively employed for the rapid

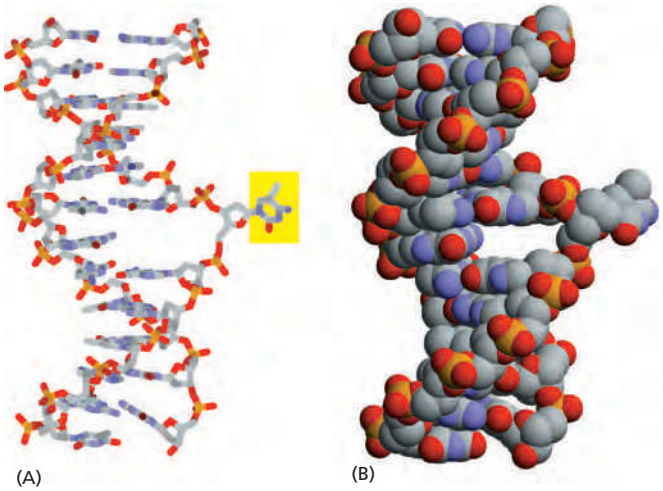


Figure 5–42 The recognition of an unusual nucleotide in DNA by base-flipping. The DNA glycosylase family of enzymes recognizes specific inappropriate bases in the conformation shown. Each of these enzymes cleaves the glycosyl bond that connects a particular recognized base *(yellow)* to the backbone sugar, removing it from the DNA. (A) Stick model; (B) space-filling model.

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.

Coupling Nucleotide Excision 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 nucleotide excision repair pathway. As discussed above, this repair system can correct many different types of DNA damage. RNA polymerase stalls at DNA lesions and, through the use of coupling proteins, directs the excision 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 eukaryotes, 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.

The importance of transcription-coupled excision repair is seen in people with Cockayne syndrome, which is caused by a defect in this coupling. 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 optimal 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 5–43**A). The addition of a second amino group to hypoxanthine

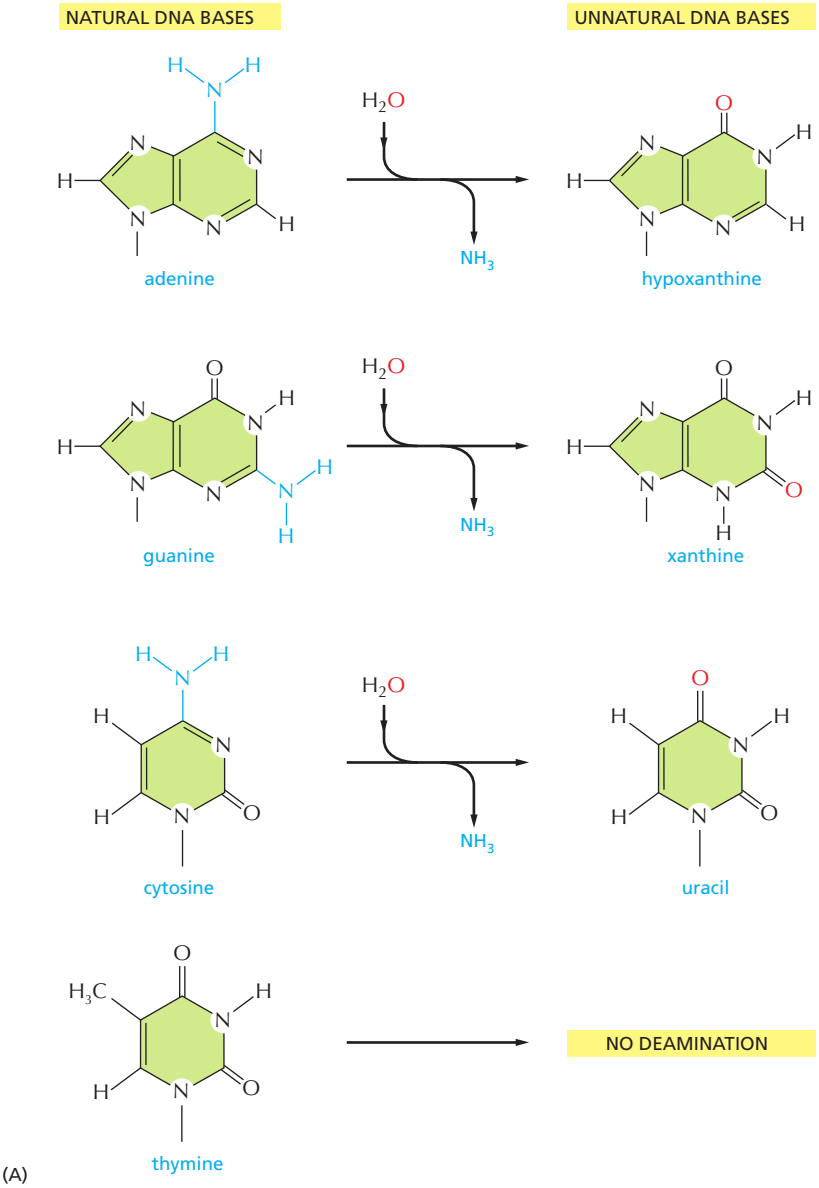


Figure 5–43 The deamination of DNA nucleotides. In each case, the oxygen atom that is added in this reaction with water is colored *red*. (A) The spontaneous deamination products of A and G are recognizable as unnatural when they occur in DNA and thus are readily found and repaired. The deamination of C to U was also illustrated in Figure 5–38; 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.

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 not be able 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 CG 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 5–43B) 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 exceptionally 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.

Special Translesion DNA Polymerases Are Used in Emergencies

If a cell's DNA suffers heavy damage, 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, backup polymerases, known as *translesion polymerases*, to replicate through the DNA damage.

Human cells have seven translesion polymerases, some of which can recognize a specific type of DNA damage and correctly 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 translesion polymerases lack exonucleolytic proofreading activity; in addition, many are much less discriminating than the replicative polymerase in choosing which nucleotide to incorporate initially. Presumably for this reason, each such translesion polymerase is given a chance to add only one or a few nucleotides before the highly accurate replicative polymerase resumes DNA synthesis.

Despite their usefulness in allowing heavily damaged DNA to be replicated, these translesion polymerases do, as noted above, pose risks to the cell. They are probably responsible for most of the base-substitution and single-nucleotide deletion mutations that accumulate in genomes; although they generally produce mutations when copying damaged DNA (see Figure 5–40), they probably also create mutations—at a low level—on undamaged DNA. Clearly, it is important for the cell to tightly regulate these polymerases, releasing them only at sites of DNA damage. Exactly how this happens for each translesion polymerase remains to be discovered, but a conceptual model is given in **Figure 5–44**. The principle of this model applies to many of the DNA repair processes discussed in this chapter: because the enzymes that carry out these reactions are potentially dangerous to the genome, they must be brought into play only at sites of damage.

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

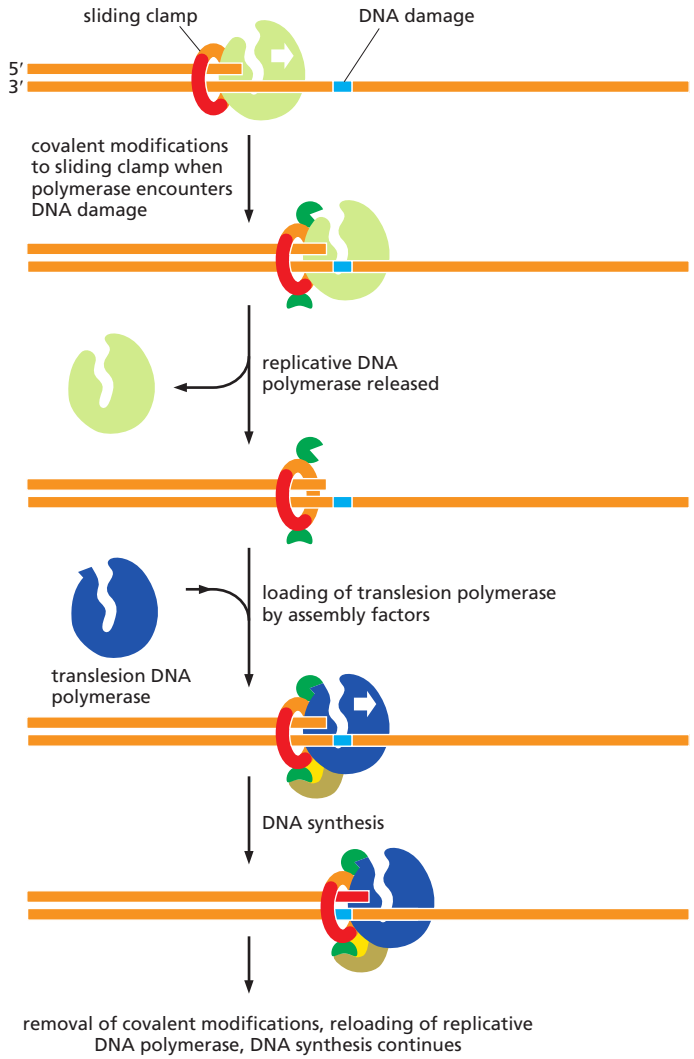


Figure 5–44 Translesion DNA polymerases can use damaged templates. According to this model, a replicative polymerase stalled at a site of DNA damage is recognized by the cell as needing rescue. Specialized enzymes covalently modify the sliding clamp (typically, it is ubiquitylated—see Figure 3–69) which releases the replicative DNA polymerase and, together with damaged DNA, attracts a translesion polymerase specific to that type of damage. Once the damaged DNA is bypassed, the covalent modification of the clamp is removed, the translesion polymerase dissociates, and the replicative polymerase is brought back into play.

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 loss of genes when the cell divides. However, two distinct mechanisms have evolved to deal with this type of damage (**Figure 5–45**). 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 nucleotides at the site of joining (**Figure 5–46**). This end-joining mechanism, which can 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 is essential for life that this mechanism is apparently an acceptable solution to the problem of rejoining broken 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. But nonhomologous end joining presents another danger: because there seems to be no mechanism to ensure that two ends being joined were originally next to each other in the genome, nonhomologous end joining can occasionally generate rearrangements in which one broken chromosome becomes covalently attached to another. This can result in chromosomes with two centromeres and chromosomes lacking centromeres altogether; both

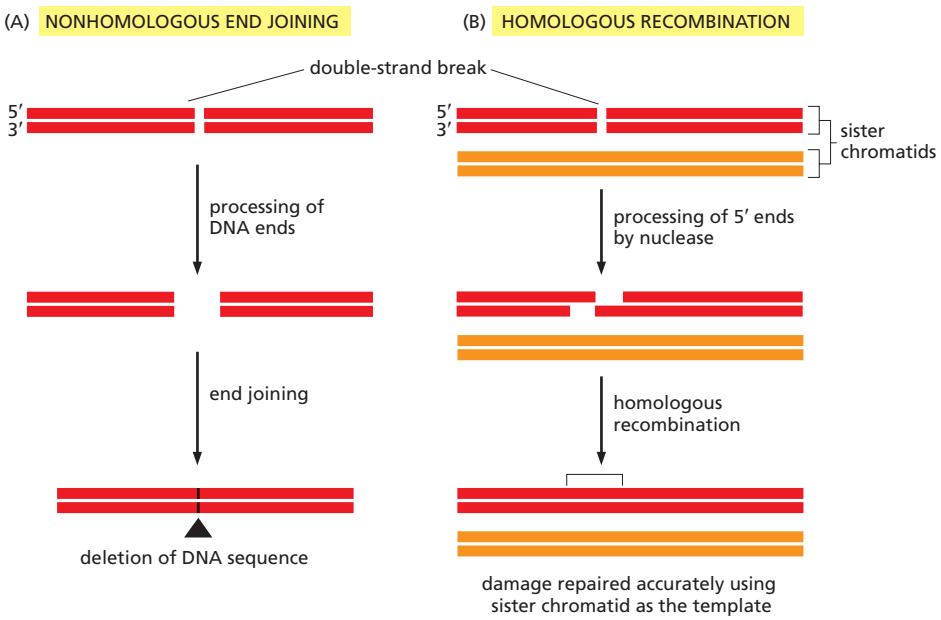


Figure 5–45 Two ways to repair double-strand breaks. (A) Nonhomologous end joining alters the original DNA sequence when repairing a broken chromosome. The initial degradation of the broken DNA ends is important because the nucleotides at the site of the initial break are often damaged and cannot be ligated. Nonhomologous end joining usually takes place when cells have not yet duplicated their DNA. (B) Repairing double-strand breaks by homologous recombination is more difficult to accomplish but restores the original DNA sequence. It typically takes place after the DNA has been duplicated (when a duplex template is available) but before the cell has divided. Details of the homologous recombination pathway are presented in the following section (see Figure 5–48).

types of aberrant chromosomes are missegregated during cell division. As previously discussed, the specialized structure of telomeres prevents the natural ends of chromosomes from being mistaken for broken DNA and "repaired" in this way.

A much more accurate type of double-strand break repair occurs in newly replicated DNA (Figure 5–45B). 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.

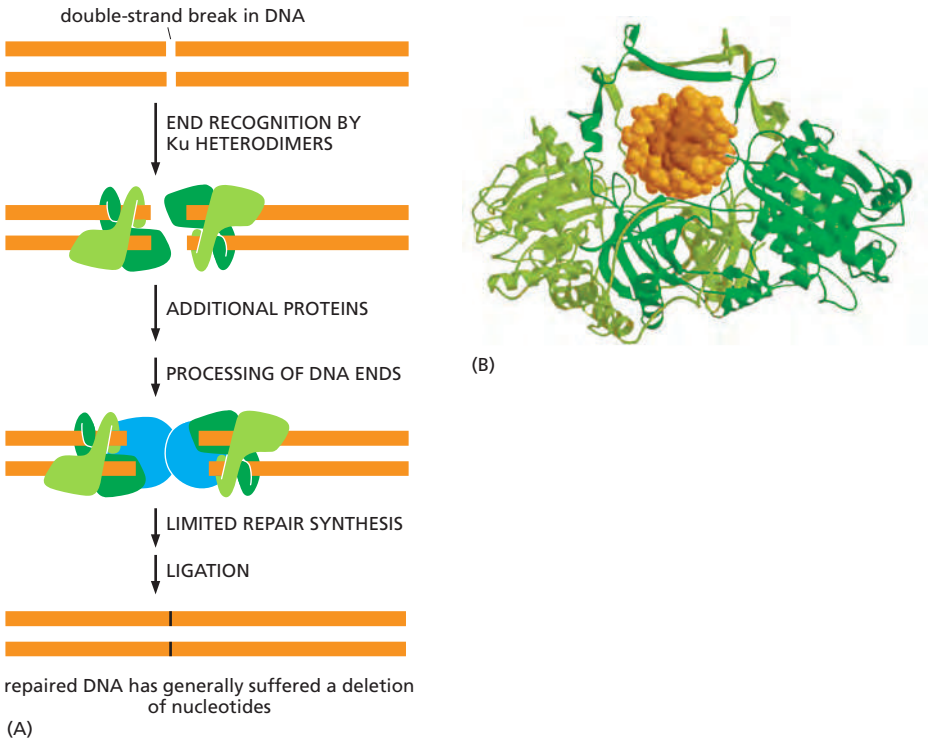


Figure 5–46 Nonhomologous end joining. (A) A central role is played by the Ku protein, a heterodimer that grasps the broken chromosome ends. The additional proteins shown are needed to hold the broken ends together while they are processed and eventually joined covalently. (B) Three-dimensional structure 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 24). 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 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 (**Movie 5.7**). Because of the importance of maintaining intact, undamaged DNA from generation to generation, eukaryotic 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 stopped if damaged DNA is detected, and it restarts when the damage has been repaired. 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 G_2 phase to M phase. These delays facilitate DNA repair by providing the time needed for the repair to reach completion.

DNA damage also results in an increased synthesis of some DNA repair enzymes. This response depends on special signaling proteins that sense DNA damage and up-regulate the appropriate DNA repair enzymes. The importance of this mechanism is revealed 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 sound the alarm in response to many types of spontaneous DNA damage (see Figure 17–62), 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. When DNA damage is excessive, a special class of inaccurate DNA polymerases, called translesion polymerases, is used to bypass the damage, allowing the cell to survive but sometimes creating permanent mutations at the sites of damage.

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, which ensures 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 the DNA repair mechanisms, a diverse set of reactions known collectively as *homologous recombination*. The key feature of **homologous recombination** (also known as *general recombination*) is an exchange of DNA strands between a pair of homologous duplex DNA

sequences, that is, segments of double helix that are very similar or identical in nucleotide sequence. This exchange allows one stretch of duplex DNA to act as a template to restore lost or damaged information on a second stretch of duplex DNA. Because the template for repair is not limited to the strand complementary to that containing the damage, homologous recombination can repair many types of DNA damage. It is, for example, the main way to accurately repair double-strand breaks, as introduced in the previous section (see Figure 5–45B). Double-strand breaks can result from radiation and reactive chemicals, but most of the time they arise from DNA replication forks that become stalled or broken independently of any such external cause. Homologous recombination accurately corrects these accidents and, because they occur during nearly every round of DNA replication, this repair mechanism is essential for every proliferating cell. Homologous recombination is 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.

Additionally, we shall see that homologous recombination plays a special role in sexually reproducing organisms. During meiosis, a key step in gamete (sperm and egg) production, it catalyzes the orderly exchange of bits of genetic information between corresponding (homologous) maternal and paternal chromosomes to create new combinations of DNA sequences in the chromosomes passed to the offspring.

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 original discovery as a key component in the specialized process of meiosis in plants and animals. The subsequent recognition that homologous recombination also occurs in 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 eukaryotes 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. Close relatives of these proteins have been found in more complex eukaryotes including flies, mice, and humans, and more recently, it has been possible to directly analyze homologous recombination in these species as well. These studies reveal that the fundamental processes that catalyze homologous recombination are common to all cells.

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 5–47**).

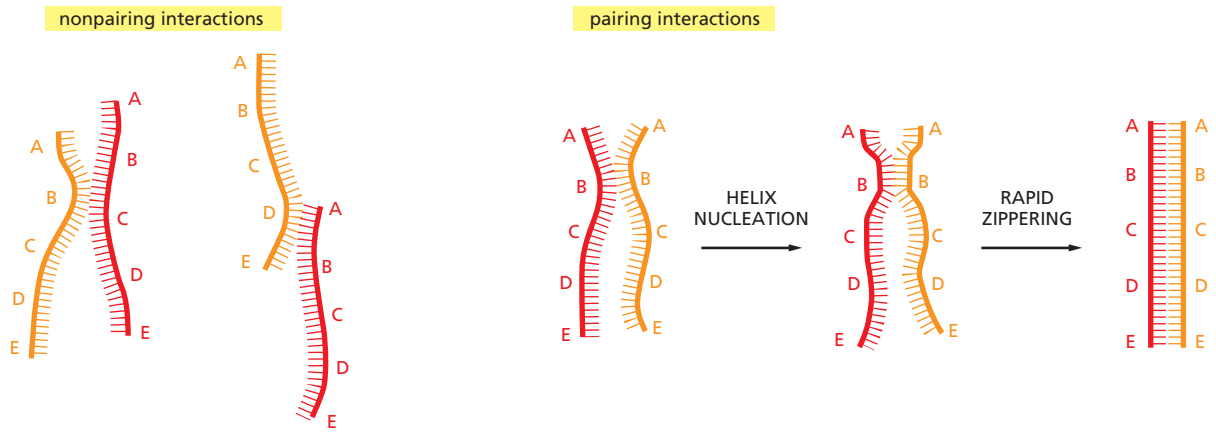


Figure 5–47 DNA hybridization. DNA double helices can re-form from their separated strands in a reaction that depends on the random collision of two complementary DNA strands. The vast majority of such collisions are not productive, as shown on the left, but a few result in a short region where complementary 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.

DNA hybridization can create a region of DNA double helix consisting of strands that originate from two different duplex DNA molecules as long as they are complementary, or nearly so. As we will see shortly, the formation of such a hybrid molecule, known as a heteroduplex, is an essential feature of homologous recombination. DNA hybridization and heteroduplex formation is also the basis for many of the methods used to study cells, and we will discuss these uses in Chapter 8.

The DNA in a living cell is almost all in the stable double-helical form, so the reaction depicted in Figure 5–47 rarely occurs *in vivo*. Instead, as we shall see, homologous recombination is brought about through a carefully controlled set of reactions that allow two DNA duplexes to sample each other's sequences without fully dissociating into single strands.

Homologous Recombination Can Flawlessly Repair Double-Strand Breaks in DNA

We saw in the previous section that nonhomologous end-joining occurs without a template and usually leaves a mutation at the site at which a double-strand break is repaired. In contrast, homologous recombination can repair double-strand breaks accurately, without any loss or alteration of nucleotides at the site of repair. For homologous recombination to do this repair job, the broken DNA has to be brought into proximity with homologous but unbroken DNA, which can serve as a template for repair. For this reason, homologous recombination often occurs just after DNA replication, when the two daughter DNA molecules lie close together and one can serve as a template for repair of the other. As we shall see, the process of DNA replication itself creates a special risk of accidents requiring this sort of repair.

The simplest pathway through which homologous recombination can repair double-strand breaks is shown in **Figure 5–48**. In essence, the broken DNA duplex and the template duplex carry out a "strand dance" so that one of the damaged strands can use the complementary strand of the intact DNA duplex as a template for repair. First, the ends of the broken DNA are chewed back, or "resected," by specialized nucleases to produce overhanging, single-strand 3′ ends. The next step is **strand exchange** (also called *strand invasion*), during which one of the single-strand 3′ ends from the damaged DNA molecule worms its way into the template duplex and searches it for homologous sequences through base-pairing. We describe this remarkable reaction in detail in the next section. Once stable base-pairing is established (which completes the strand exchange step), an accurate DNA polymerase extends the invading strand by using the information provided by the undamaged template molecule, thus restoring the damaged DNA. The last steps—strand displacement, further repair synthesis, and ligation—restore the two original DNA double helices and complete the repair process. Homologous recombination resembles other DNA repair reactions in that a

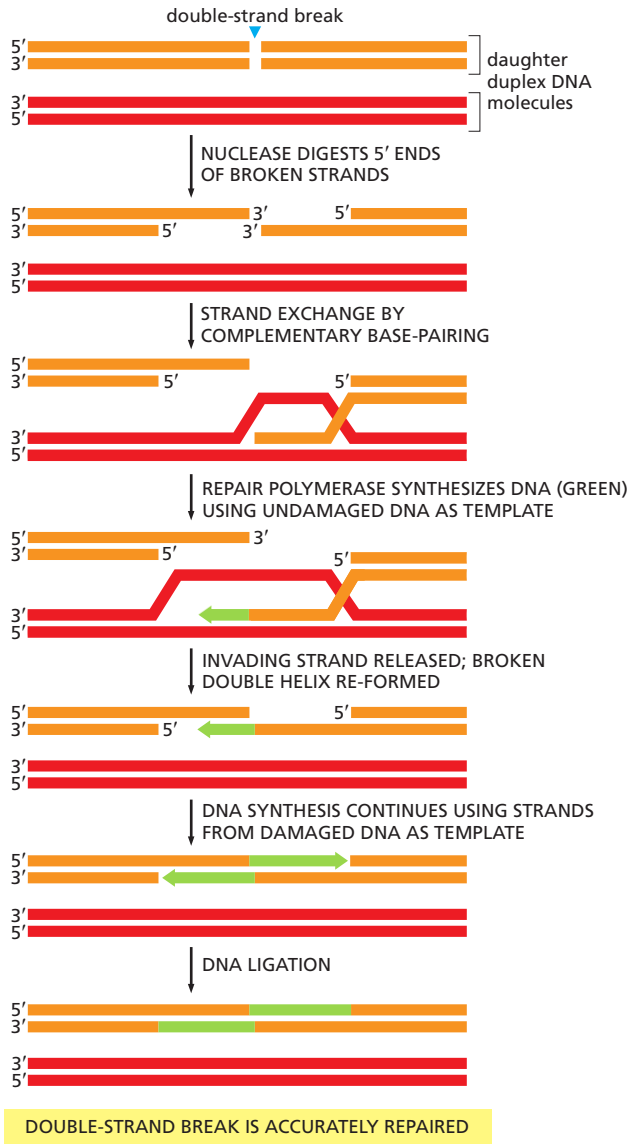


Figure 5–48 Mechanism of double-strand break repair by homologous recombination. This is the preferred method for repairing DNA double-strand breaks that arise shortly after the DNA has been replicated, while the daughter DNA molecules are still held close together. In general, homologous recombination 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 and new DNA synthesis, indicated in *green*.

DNA polymerase utilizes a pristine template to restore damaged DNA. However, instead of using the partner complementary strand as a template, as occurs in most DNA repair pathways, homologous recombination exploits a complementary strand from a separate DNA duplex.

Strand Exchange Is Carried Out by the RecA/Rad51 Protein

Of all the steps of homologous recombination, strand exchange is the most difficult to imagine. How does the invading single strand rapidly sample a DNA duplex for homology? Once the homology is found, how does the exchange occur? How is the inherent stability of the template double helix overcome?

The answers to these questions came from biochemical and structural studies of the protein that carries out these feats, called **RecA** in *E. coli* and **Rad51** in virtually all eukaryotic organisms. To catalyze strand exchange, RecA first binds cooperatively to the invading single strand, forming a protein–DNA filament that forces the DNA into an unusual configuration: groups of three consecutive nucleotides are held as though they were in a conventional DNA double helix but, between adjacent triplets, the DNA backbone is untwisted and stretched out (**Figure 5–49**). This unusual protein–DNA filament then binds to duplex DNA

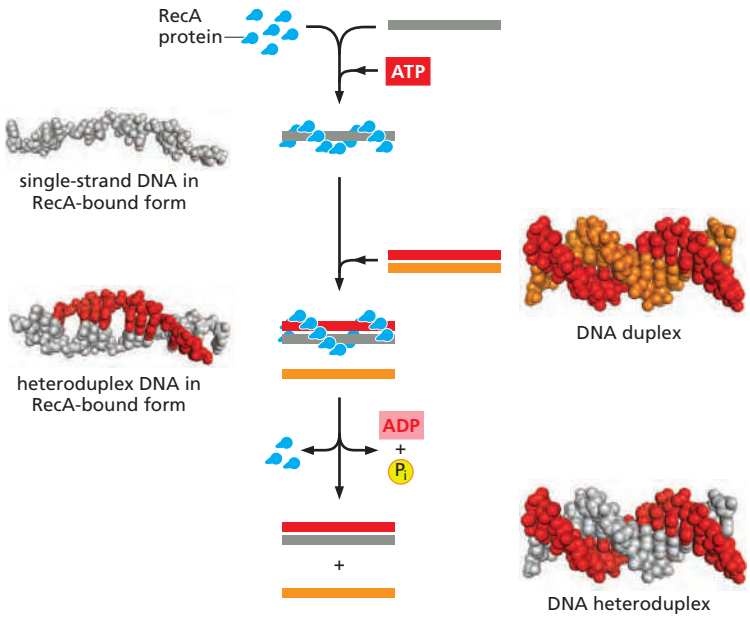


Figure 5–49 Strand invasion catalyzed by the RecA protein. Our understanding of this reaction is based in part on structures determined by x-ray diffraction studies of RecA bound to single- and double-strand DNA. These DNA structures (shown without the RecA protein) are on the left side of the diagram. Starting at the top, ATP-bound RecA associates with single-strand DNA, holding it in an elongated form where groups of three bases are separated from each other by a stretched and twisted backbone. In the next step, the RecA-bound single strand then binds to duplex DNA, destabilizing it and allowing the single strand to sample its sequence through base-pairing, three bases at a time. If no match is found, the RecA-bound single strand of DNA rapidly dissociates and begins a new search. If an extensive match is found, the structure is disassembled through ATP hydrolysis, resulting in the dissociation of RecA and the exchange of one single strand of DNA for another, thereby forming a heteroduplex. (PDB code: 3CMX.)

in a way that stretches the duplex, destabilizing it and making it easy to pull the strands apart. The invading single strand then can sample the sequence of the duplex by conventional base-pairing. This sampling occurs in triplet nucleotide blocks: if a triplet match is found, the adjacent triplet is sampled, and so on. In this way, mismatches quickly lead to dissociation and only an extended stretch of base-pairing (at least 15 nucleotides) stabilizes the invading strand and leads to strand exchange.

RecA hydrolyzes ATP, and the steps described above require that each RecA monomer along the filament be in the ATP-bound state. However, the searching itself does not require ATP hydrolysis; instead, the process occurs by simple molecular collision, allowing many potential sequences to be rapidly sampled. Once the strand-exchange reaction is completed, however, ATP hydrolysis is necessary to disassemble RecA from the complex of DNA molecules. At this point, repair DNA polymerases and DNA ligase can complete the repair process, as shown in Figure 5–48.

Homologous Recombination Can Rescue Broken DNA Replication Forks

Although accurately repairing double-strand breaks, which can arise from radiation or chemical reactions, is a crucial function of homologous recombination, perhaps its most important role is in rescuing stalled or broken DNA replication forks. Many types of events can cause a replication fork to break, and here we 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. The broken fork can be flawlessly repaired (**Figure 5–50**) using the same basic homologous recombination reactions we discussed above for the repair of double-strand breaks. With slight modifications, the set of reactions depicted in Figures 5–48 and 5–50—known collectively as homologous recombination—can accurately repair many different types of DNA damage.

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

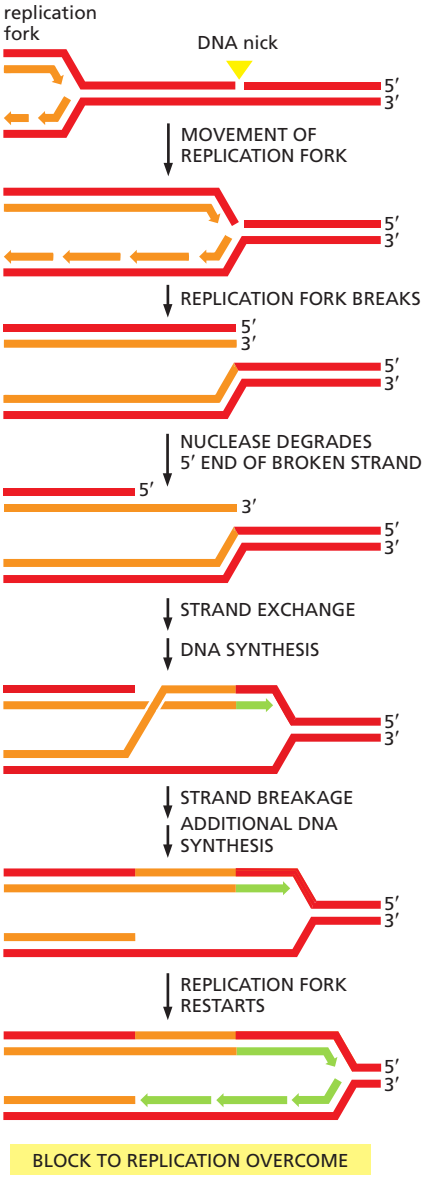


Figure 5–50 Repair of a broken replication fork by homologous recombination. When a moving replication fork encounters a single-strand break, it will collapse, but can be repaired by homologous recombination. The process uses many of the same reactions shown in Figure 5–48 and proceeds through the same basic steps. *Green* strands represent the new DNA synthesis that takes place after the replication fork has broken. This pathway allows the fork to move past the site that was nicked on the original template by using the undamaged duplex as a template to synthesize DNA. (Adapted from M.M. Cox, *Proc. Natl Acad. Sci. USA* 98:8173–8180, 2001. With permission from National Academy of Sciences.)

some dangers to the cell as it sometimes "repairs" damage using the wrong bit of the genome as the template. For example, sometimes a broken human chromosome is "repaired" using the homolog from the other parent instead of the sister chromatid as the template. Because maternal and paternal chromosomes differ in DNA sequence at many positions along their lengths, this type of repair can convert the sequence of the repaired DNA from the maternal to the paternal sequence or vice versa. The result of this type of errant recombination is known as **loss of heterozygosity**. It can have severe consequences if the homolog used for repair contains a deleterious mutation, because the recombination event destroys the "good" copy. Loss of heterozygosity, although rare, is a critical step in the formation of many cancers (discussed in Chapter 20).

Cells go to great lengths to minimize the risk of mishaps of these types; indeed, nearly every step of homologous recombination is carefully regulated. For example, the first step, processing of the broken ends, is coordinated with the cell cycle: the nuclease enzymes that carry out this process are activated (in part, by phosphorylation) only in the S and G_2 phases of the cell cycle, when a daughter duplex (either as a partially replicated chromosome or a fully replicated sister chromatid) can serve as a template for repair (see Figure 5–50). The close proximity of the two daughter chromosomes disfavors the use of other genome sequences in the repair process.

The loading of RecA or Rad52 onto the processed DNA ends and the subsequent strand-exchange reaction are also tightly controlled. Although these proteins alone can carry out these steps *in vitro*, a series of accessory proteins, including Rad52, is needed in eukaryotic cells to ensure that homologous recombination is efficient and accurate (**Figure 5–51**). There are many such accessory proteins, and exactly how they coordinate and control homologous recombination remains a mystery. We do know that the enzymes that catalyze recombinational repair are made at relatively high levels in eukaryotes and are dispersed throughout the nucleus in an inactive form. In response to DNA damage, they rapidly converge on the sites of DNA damage, become activated, and form "repair factories" where many lesions are apparently brought together and repaired (**Figure 5–52**).

In Chapter 20, we shall see that both too much and too little homologous recombination can lead to cancer in humans, the former through repair using the "wrong" template (as described above) 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.

Not surprisingly, mutations in the components that carry out and regulate homologous recombination are responsible for several inherited forms of cancer. Two of these, the Brca1 and Brca2 proteins, were first discovered because

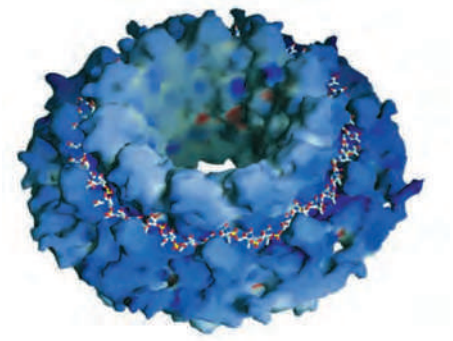


Figure 5–51 Structure of a portion of the Rad52 protein. This doughnut-shaped structure is composed of 11 subunits. Single-strand DNA has been modeled into the deep groove running along the protein surface. Rad52 helps load Rad51 onto single-strand DNA to form the nucleoprotein filament that carries out strand exchange. Rad52 also acts later to re-form the double helix and complete the homologous recombination reaction. (From M.R. Singleton et al., *Proc. Natl Acad. Sci. USA* 99:13492–13497, 2002. With permission from National Academy of Sciences.)

mutations in their genes lead to a greatly increased frequency of breast cancer. Because these mutations cause inefficient repair by homologous recombination, accumulation of DNA damage can, in a small proportion of cells, give rise to a cancer. Brca1 regulates an early step in broken-end processing; without it, such ends are not processed correctly for homologous recombination and instead are repaired inaccurately by the nonhomologous end-joining pathway (see Figure 5–45). Brca2 binds to the Rad51 protein, preventing its polymerization on DNA, and thereby maintaining it in an inactive form until it is needed. Normally, upon DNA damage, Brca2 helps to bring Rad51 protein rapidly to sites of damage and, once in place, to release it in its active form onto single-strand DNA.

Homologous Recombination Is Crucial for Meiosis

We have seen that homologous recombination comprises a group of reactions—including broken-end processing, strand exchange, limited DNA synthesis, and ligation—to exchange DNA sequences 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 that carry novel combinations of genes as a result of the deliberate exchange of material between different chromosomes. Although this occasionally occurs by accident in mitotic cells (and is often detrimental), it is a frequent and necessary part of meiosis, which occurs in sexually reproducing organisms such as fungi, plants, and animals.

Here, homologous recombination occurs as an integral part of the process whereby chromosomes are parceled out to germ cells (sperm and eggs in animals). We discuss the process of meiosis in detail in Chapter 17; in the following sections, we discuss how homologous recombination during meiosis produces chromosome *crossing-over* and *gene conversion,* resulting in hybrid chromosomes that contain genetic information from both the maternal and paternal homologs (**Figure 5–53**). Crossing-over and gene conversion are both generated by homologous recombination mechanisms that, at their core, resemble those used to repair double-strand breaks.

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 (**Figure 5–54**). Like a topoisomerase, Spo11, after catalyzing this reaction, remains covalently bound to the broken

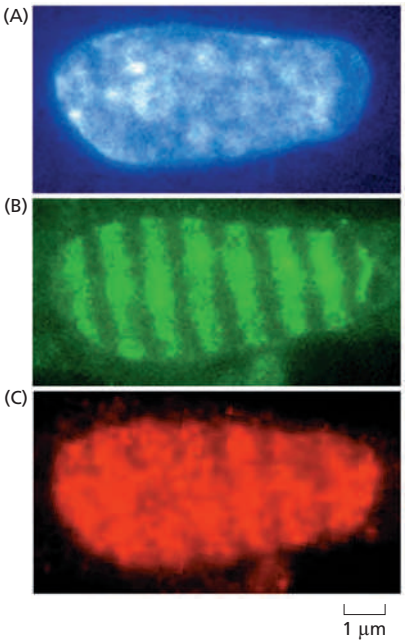


Figure 5–52 Experiment demonstrating the rapid localization of repair proteins to DNA double-strand breaks. Human fibroblasts were x-irradiated to produce DNA double-strand breaks. Before the x-rays struck the cells, they were passed through a microscopic grid with x-ray-absorbing "bars" spaced 1 μm 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 due to repair of DNA damage, indicated by incorporation of BudR (a thymidine analog) and subsequent staining with fluorescently labeled antibodies to BudR *(green).* (C) Localization of the Mre11 complex to damaged DNA as visualized by antibodies against the Mre11 subunit *(red).* Mre11 is a nuclease that processes damaged DNA in preparation for homologous recombination (see Figure 5–48). (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.)

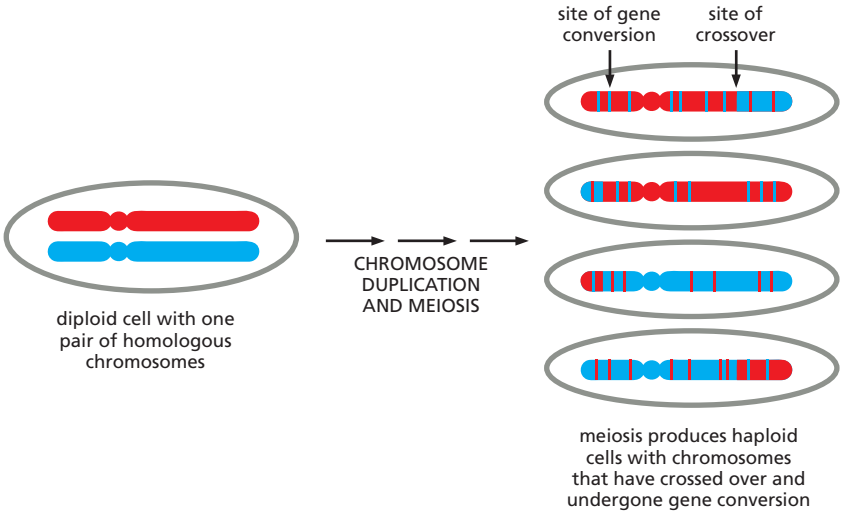


Figure 5–53 Chromosome crossing-over occurs in meiosis. Meiosis is the process by which a diploid cell gives rise to four haploid germ cells, as described in detail in Chapter 17. Meiosis produces germ cells in which the paternal and maternal genetic information (*red* and *blue*) has been reassorted through chromosome crossovers. In addition, many short regions of gene conversion occur, as indicated.

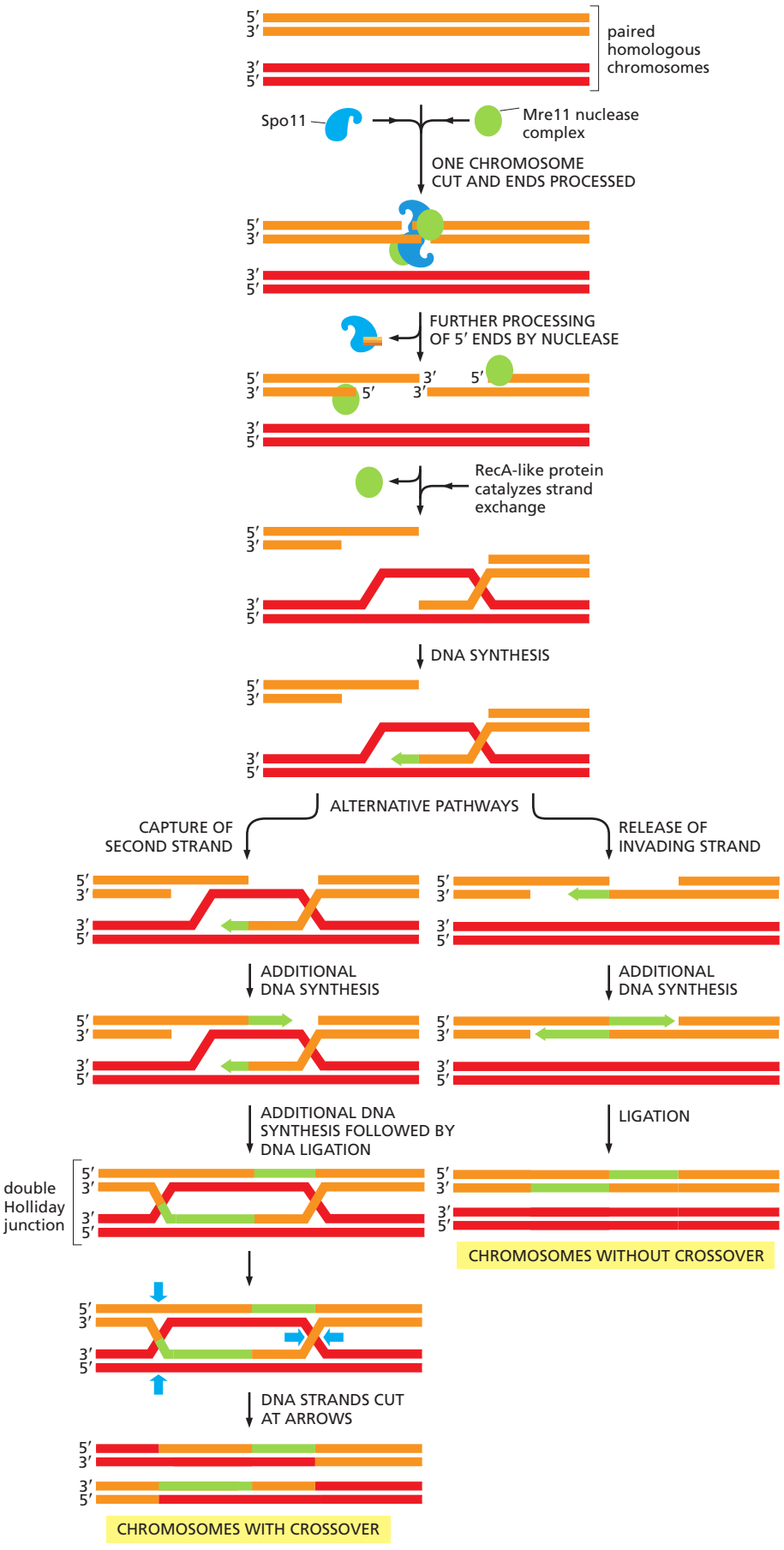


Figure 5–54 Homologous recombination during meiosis can generate chromosome crossovers. Once the meiosis-specific protein Spo11 and the Mre11 complex break the duplex DNA and process the ends, homologous recombination can proceed along alternative pathways. One (right side of figure) closely resembles the double-strand break repair reaction shown in Figure 5–48 and results in chromosomes that have been "repaired" but have not crossed over. The other (left side with strand breaks as shown by the *blue arrows)* proceeds through a double Holliday junction and produces two chromosomes that have crossed over. During meiosis, homologous recombination takes place between maternal and paternal chromosome homologs when they are held tightly together (see Figure 17–54).

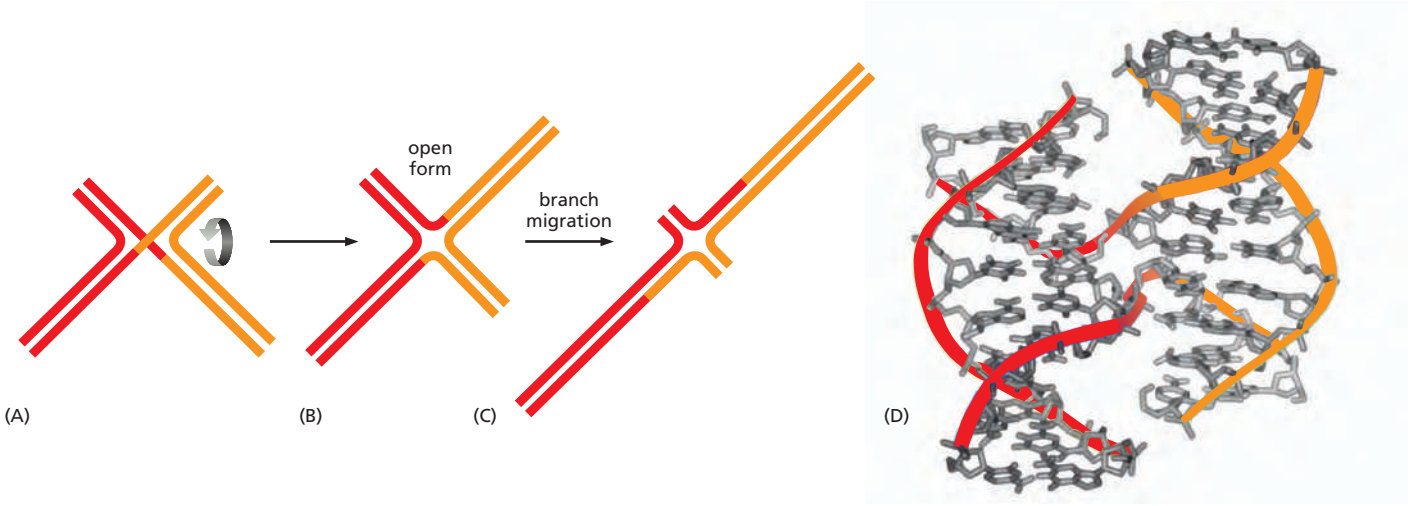


Figure 5–55 A Holliday junction. The initially formed structure (A) is usually drawn with two strands crossing, as in Figure 5–54. An isomerization of the Holliday junction (B) produces an open, symmetrical structure that is bound by specialized proteins. (C) These proteins "move" the Holliday junctions by a coordinated set of branch-migration reactions (see Figure 5–57 and Movie 5.8). (D) Structure of the Holliday junction in the open form depicted in (B). The Holliday junction is named for the scientist who first proposed its formation. (PDB code: 1DCW.)

DNA (see Figure 5–21). A specialized nuclease then rapidly degrades the ends bound by Spo11, removing the protein along with the DNA and leaving protruding 3′ single-strand ends.

At this point, many of the recombination reactions resemble those described above for the repair of double-strand breaks; indeed, some of the same proteins are used for both processes. However, several meiosis-specific proteins direct them to perform their tasks somewhat differently, resulting in the distinctive outcomes observed for meiosis. 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. In the sections that follow, we describe in more detail those aspects of homologous recombination that are especially important for meiosis.

Holliday Junctions Are Formed During Meiosis

Of special importance in meiosis is an intermediate known as a **Holliday junction** or *cross-strand exchange* (Figure 5–55). Each Holliday junction can adopt multiple conformations and a special set of recombination proteins binds to, and thereby stabilizes, the open, symmetric isomer.

Specialized proteins that bind to Holliday junctions can catalyze a reaction known as *branch migration* (Figure 5–56), whereby DNA is spooled through the Holliday junction by continually breaking and re-forming base pairs (Figure 5–57). In this way, the Holliday junction proteins use ATP hydrolysis to expand the region of heteroduplex DNA initially created by the strand-exchange reaction. In meiosis, heteroduplex regions often "migrate" thousands of nucleotides from the original site of the double-strand break. As shown in Figure 5–54, Holliday junctions usually occur in pairs, known as double Holliday junctions.

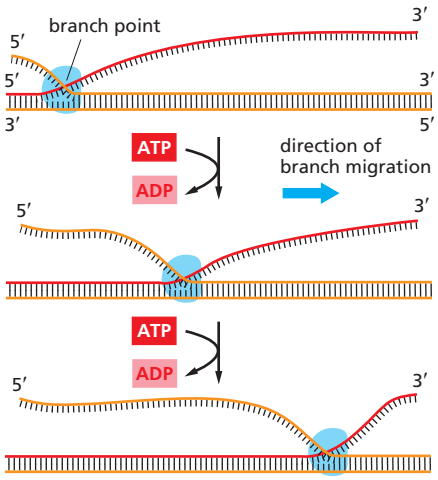


Figure 5–56 Simplified view of branch migration. In branch migration, base pairs are continually broken and formed as the branch point moves. Although branch migration can happen spontaneously on naked DNA molecules, the process is inefficient and the branch moves back and forth at random. In the cell, branch migration is carried out using specialized proteins and ATP hydrolysis to ensure that, as shown, the branch moves rapidly and in one direction. As shown in Figure 5–57, branch migrations often occur at Holliday junctions, where two branch-migration reactions are coupled.

Homologous Recombination Produces Both Crossovers and Non-Crossovers During Meiosis

As shown in Figure 5–54, there are two basic outcomes of homologous recombination during meiosis. In humans, approximately 90% of the double-strand breaks produced during meiosis are resolved as non-crossovers (see right side of Figure 5–54). Here, the two original DNA duplexes separate from each other in a form unaltered except for a region of heteroduplex that formed near the site of the original double-strand break. This set of reactions resembles that described above for the repair of double-strand breaks (see Figure 5–48).

The other outcome is more profound: a double Holliday junction is formed and is cleaved by specialized enzymes to create a *crossover* (see left side of Figure 5–54). The two original portions of each chromosome upstream and downstream

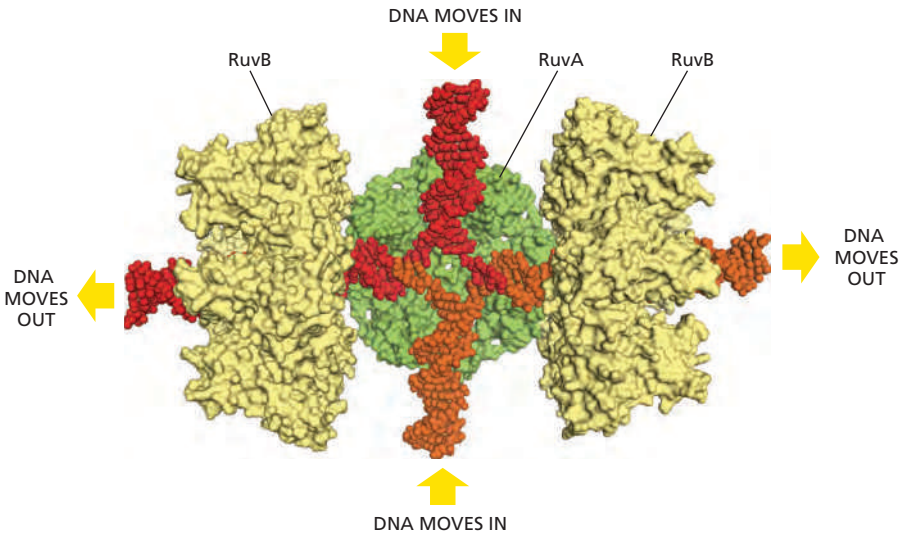


Figure 5–57 Enzyme-catalyzed branch movement at a Holliday junction by branch migration. In *E. coli*, a tetramer of the RuvA protein *(green)* and two hexamers of the RuvB protein *(yellow)* bind to the open form of the junction. The RuvB protein, which resembles the hexameric helicases used in DNA replication (Figure 5–14), uses the energy of ATP hydrolysis to spool DNA rapidly through the Holliday junction, extending the heteroduplex region as shown. The RuvA protein coordinates this movement, threading the DNA strands to avoid tangling. (PDB codes: 1IXR, 1C7Y.)

from the two Holliday junctions are thereby swapped, creating two chromosomes that have crossed over.

How does the cell decide which Spo11-induced double-strand breaks to resolve as crossovers? The answer is not yet known, but we know the decision is an important one. The relatively few crossovers that do form are distributed along chromosomes in such a way that a crossover in one position inhibits crossing-over in neighboring regions. Termed *crossover control*, this fascinating but poorly understood regulatory mechanism ensures the roughly even distribution of crossover points along chromosomes. It also ensures that each chromosome—no matter how small—undergoes at least one crossover every meiosis. For many organisms, roughly two crossovers per chromosome occur during each meiosis, one on each arm. As discussed in detail in Chapter 17, 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 with the DNA sequence of the paternal homolog is base-paired with a strand from the maternal homolog (**Figure 5–58**). These heteroduplex regions can tolerate a small percentage of mismatched base pairs, and because of branch migration, they often extend for thousands of nucleotide pairs. The many non-crossover events that occur in meiosis thereby produce scattered sites in the germ cells where short DNA sequences from one homolog have been pasted into the other homolog. Heteroduplex regions mark sites of potential *gene conversion*—where the four haploid chromosomes produced by meiosis contain three copies of a DNA sequence from one homolog and only one copy of this sequence from the other homolog (see Figure 5–53), as explained next.

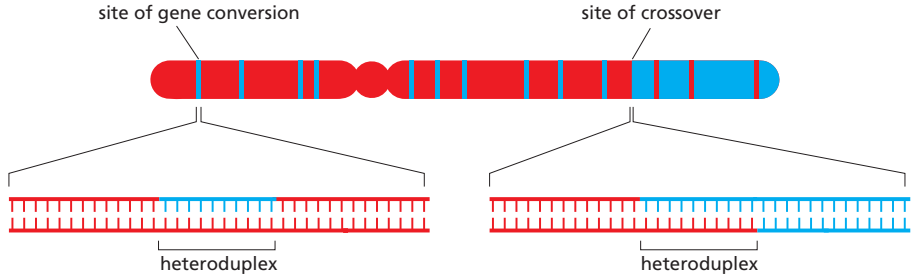


Figure 5–58 Heteroduplexes formed during meiosis. Heteroduplex DNA is present at sites of recombination that are resolved either as crossovers or non-crossovers. Because the DNA sequences of maternal and paternal chromosomes differ at many positions along their lengths, heteroduplexes often contain a small number of base-pair mismatches.

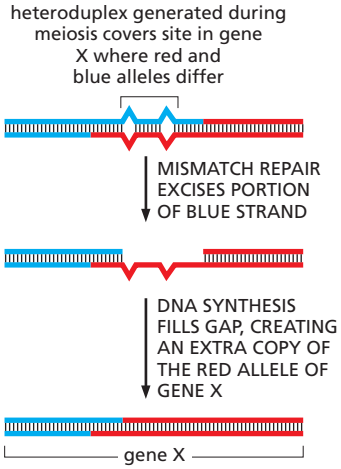


Figure 5–59 Gene conversion caused by mismatch correction. In this process, heteroduplex DNA is formed at the sites of homologous recombination between maternal and paternal chromosomes. If the maternal and paternal DNA sequences are slightly different, the heteroduplex region will include some mismatched base pairs, which may then be corrected by the DNA mismatch repair machinery (see Figure 5–19). Such repair can "erase" nucleotide sequences 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.

Homologous Recombination Often Results in Gene Conversion

In sexually reproducing organisms, it is a fundamental law of genetics that—aside from mitochondrial DNA, which is inherited only through the mother—each parent makes an equal genetic contribution to an offspring. One complete set of nuclear genes is inherited from the father and one complete set is inherited from the mother. Underlying this law is the accurate parceling out of chromosomes to the germ cells (eggs and sperm) that takes place during meiosis. Thus, when a diploid cell in a parent undergoes meiosis to produce four haploid germ cells, exactly half of the genes distributed among these four cells should be maternal (genes inherited from the mother of this parent) and the other half paternal (genes inherited from the father of this parent). 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 genetic rules. Occasionally, for example, meiosis yields three copies of the maternal version of a gene and only one copy of the paternal allele. Alternative versions of the same gene are called **alleles**, and it is the divergence from their expected distribution during meiosis that 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, but one of the most important arises from a particular consequence of recombination during meiosis. We have seen that both crossovers and non-crossovers produce heteroduplex regions of DNA. If the two strands that make up a heteroduplex region do not have identical nucleotide sequences, mismatched base pairs are formed, and these are often repaired by the cell's mismatch repair system (see Figure 5–19). However, the mismatch repair system cannot distinguish between the paternal and maternal strands and will randomly choose the strand to be used as a template. As a consequence, one allele will be lost and the other duplicated (**Figure 5–59**), 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.

Summary

Homologous recombination describes a flexible set of reactions resulting in the exchange of DNA sequences between a pair of identical or nearly identical duplex DNA molecules. In all cells, this process is essential for the error-free repair of chromosome damage, particularly double-strand breaks and broken or stalled replication forks. Homologous recombination is also responsible for the crossing-over of chromosomes that occurs during meiosis. Homologous recombination takes place through a variety of pathways, but they have in common a strand-exchange step whereby a single strand from one DNA duplex invades a second duplex and base-pairs with one strand while displacing the other. This reaction, catalyzed by the RecA/Rad51 family of proteins, can only occur if the invading strand can form a short stretch of consecutive nucleotide pairs with one of the strands of the duplex. This requirement ensures that homologous recombination occurs only between identical or very similar DNA sequences.

When used as a repair mechanism, homologous recombination occurs between a damaged DNA molecule and its recently duplicated sister molecule, with the undamaged duplex acting as a template to repair the damaged copy flawlessly.

In meiosis, homologous recombination is initiated by deliberate, carefully regulated double-strand breaks and occurs preferentially between the homologous chromosomes rather than the newly replicated sister chromatids. The outcome can be either 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.

TRANSPOSITION AND CONSERVATIVE SITE-SPECIFIC RECOMBINATION

We have seen that homologous recombination can result in the exchange of DNA sequences between chromosomes. However, the order of genes on the interacting chromosomes 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 reactions can alter gene order along a chromosome and can cause unusual types of mutations that introduce whole blocks of DNA sequence into the genome.

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 time scales, 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 4–62). 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 to our 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 many of these mutations will be deleterious to the organism, some will be advantageous and may 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 known examples of this type of movement.

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 specific DNA sequences 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 a genome. In particular, there is no general requirement for sequence similarity 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. More frequent movement would probably destroy the host cell's genome.

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*. The differences among them are briefly outlined in **Table 5–4**, and each class will be discussed in turn.

DNA-Only Transposons Can Move by a Cut-and-Paste Mechanism

DNA-only transposons, so named because they exist only as DNA during their movement, 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. Now, the situation is different—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

TABLE 5–4 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 whose production is driven 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 synthesized 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. Some viruses can also move in and out of host-cell chromosomes by transpositional mechanisms. These viruses are related to the first two classes of transposons.

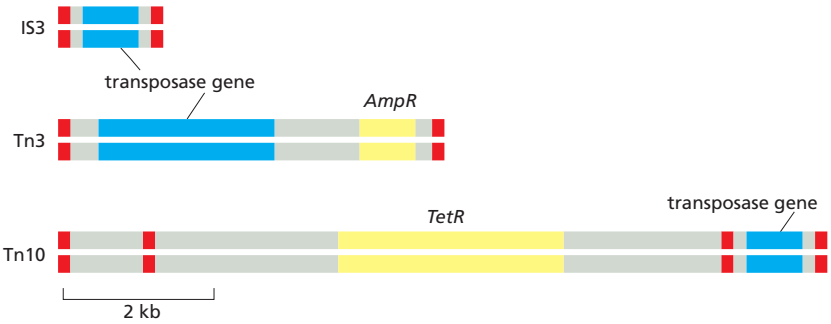


Figure 5–60 Three of the many DNA-only transposons found in bacteria. Each of these mobile DNA elements contains a gene that encodes a *transposase*, an enzyme that carries out the DNA breakage and joining reactions needed for the element to move. Each transposon also carries short DNA sequences (indicated 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)*—a penicillin derivative—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.

that encode antibiotic-inactivating enzymes that are carried on transposons (Figure 5–60). 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 (see Figure 1–19). Once introduced into a new cell, a transposon can insert itself into the genome and be faithfully passed on to all progeny cells through the normal processes of DNA replication and cell division.

DNA-only transposons can relocate from a donor site to a target site by *cut-and-paste transposition* (Figure 5–61). Here, the transposon is literally excised from one spot on a genome and inserted into another. This reaction produces a short duplication of the target DNA sequence at the insertion site; these direct repeat sequences that flank the transposon serve as convenient records of prior transposition events. Such "signatures" often provide valuable clues in identifying transposons in genome sequences.

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 5–48), provided that the chromosome has just been replicated and an identical copy of the damaged host sequence is available. Alternatively, 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 5–45).

Remarkably, the same mechanism used to excise cut-and-paste transposons from DNA has been found to operate in developing immune systems of

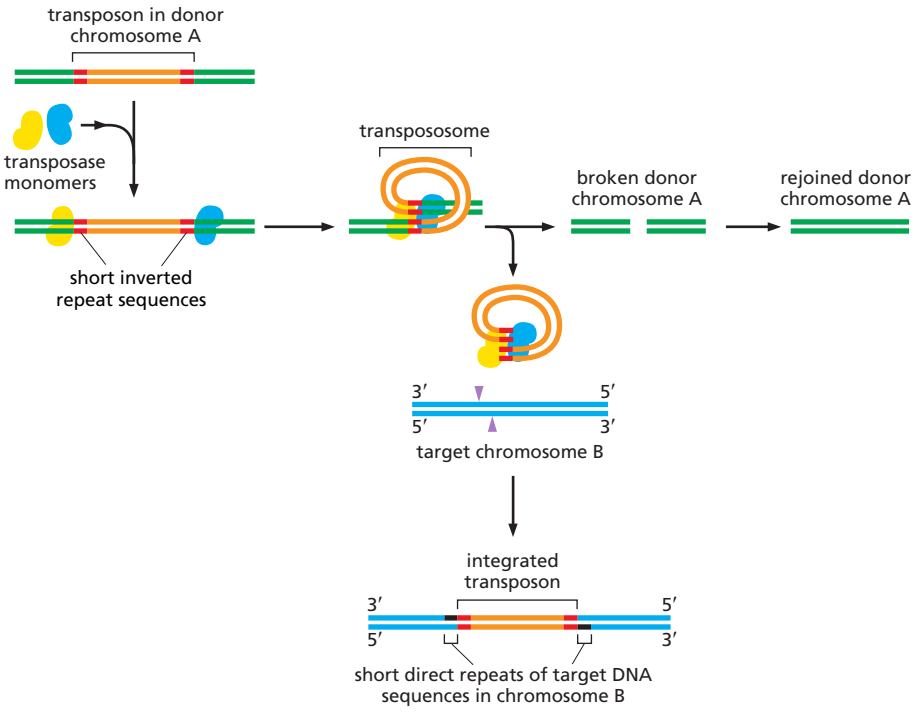


Figure 5–61 Cut-and-paste transposition. DNA-only transposons can be recognized in chromosomes by the "inverted repeat DNA sequences" *(red)* present at their ends. These sequences, which can be as short as 20 nucleotides, are all that is necessary for the DNA between them to be transposed by the particular 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, also catalyzed by the transposase, occurs at a random site through the creation of staggered breaks in the target chromosome *(purple arrowheads)*. Following the transposition reaction, the single-strand gaps created by the staggered breaks are repaired by DNA polymerase and ligase *(black)*. 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).

vertebrates, catalyzing the DNA rearrangements that produce antibody and T cell receptor diversity. Known as *V(D)J recombination*, this process will be discussed in Chapter 24. Found only in vertebrates, V(D)J recombination is a relatively recent evolutionary novelty, but it is believed to be derived from the much more ancient cut-and-paste transposons.

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 viruses that insert themselves into a host chromosome do so by employing one of the first two mechanisms listed in Table 5–4; namely, by behaving like DNA-only transposons or like retroviral-like retrotransposons. Indeed, much of our knowledge of these mechanisms has come from studies of particular viruses that employ them.

Transposition has a key role in the life cycle of many viruses. Most notable are the **retroviruses**, which include the human AIDS virus, HIV. Outside the cell, a retrovirus exists as a single-strand RNA genome packed into a protein shell or *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-strand DNA molecule by the action of this crucial enzyme, which is able to polymerize DNA on either an RNA or a DNA template (**Figure 5–62**). 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 1–4).

Once the reverse transcriptase has produced a double-strand DNA molecule, specific sequences near its two ends are recognized by a virus-encoded

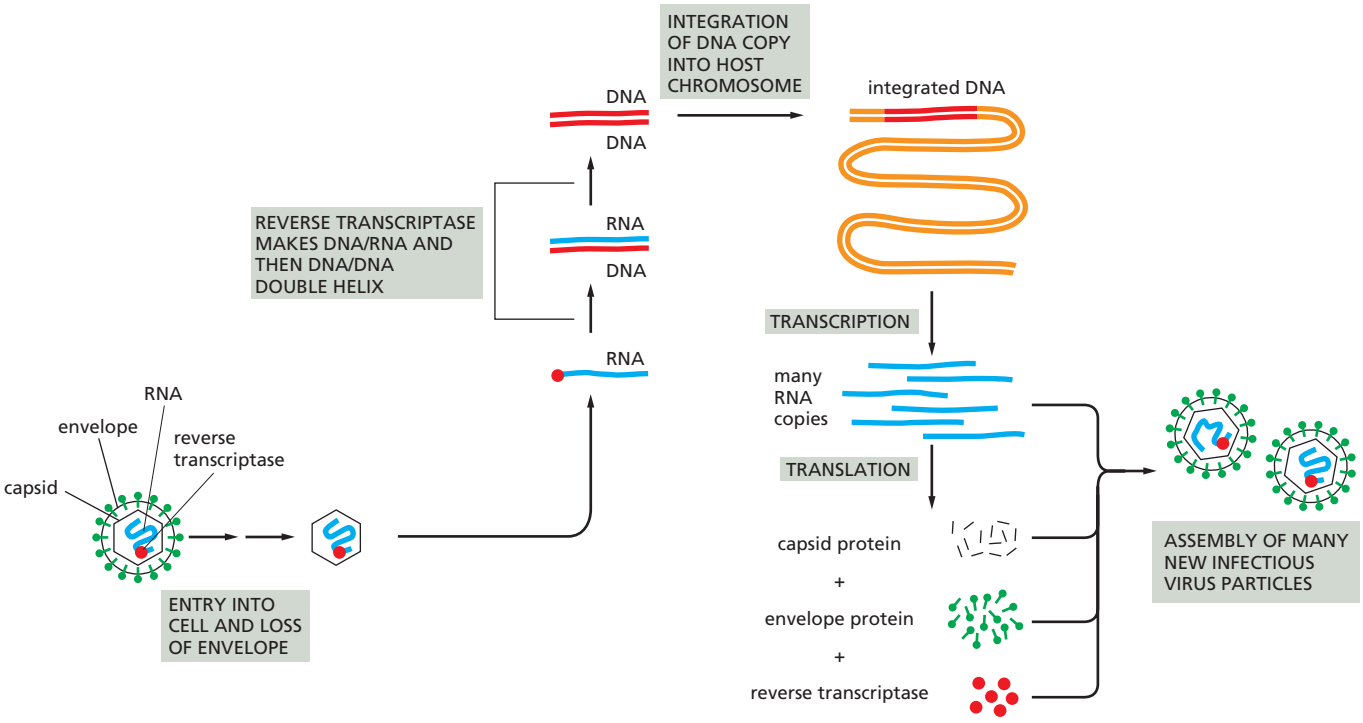


Figure 5–62 The life cycle of a retrovirus. The retrovirus genome consists of an RNA molecule *(blue)* that is typically between 7000 and 12,000 nucleotides in length. It is packaged inside a protein capsid, which is surrounded by a lipid-based envelope that contains virus-encoded envelope proteins *(green)*. Inside an infected cell, the enzyme reverse transcriptase *(red circle)* first makes a DNA copy of the viral RNA molecule and then a second DNA strand, generating a double-strand 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. 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).

transposase called *integrase*. Integrase then inserts the viral DNA into the chromosome by a mechanism similar to that used by the cut-and-paste DNA-only transposons (see Figure 5–61).

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

A large family of transposons called **retroviral-like retrotransposons** (see Table 5–4) move themselves in and out of chromosomes by a mechanism that is similar 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 processes 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-strand DNA copy of the RNA molecule via an RNA–DNA hybrid intermediate, precisely mirroring the early stages of infection by a retrovirus (see Figure 5–62). Like a retrovirus, the linear, double-strand DNA molecule then integrates into a site on the chromosome using an integrase enzyme that is also encoded by the element. The structure and mechanisms of these integrases closely resemble those of the transposases of DNA-only transposons.

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 5–4). Although most of these transposons in the human genome 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 6–24).

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 5–63**, 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 reveals 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. Together the LINEs and SINEs make up over 30% of the human genome (see Figure 4–62); there are 500,000 copies of the former and over a million of the latter.

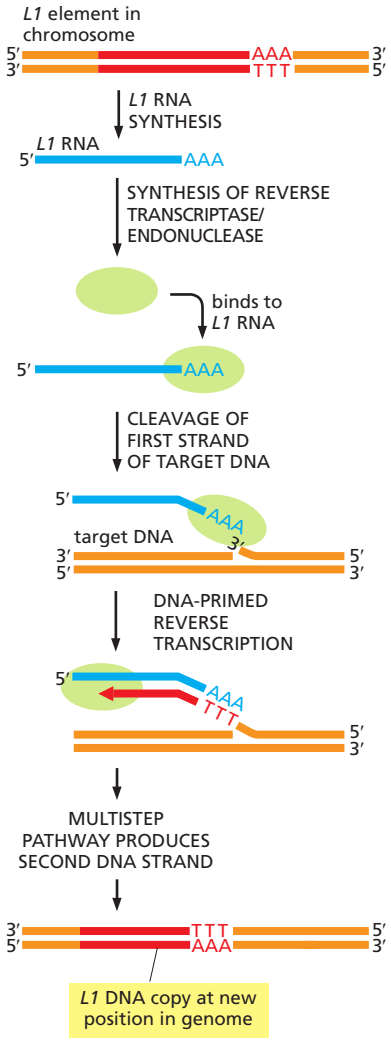


Figure 5–63 Transposition by a nonretroviral retrotransposon. Transposition of the *L1* element *(red)* begins when an endonuclease attached to the *L1* reverse transcriptase *(green)* and the *L1* RNA *(blue)* nick the target DNA at the point at which insertion will occur. This cleavage releases a 3′-OH DNA end in the target DNA, which is then used as a primer for the reverse transcription step shown. This generates a single-strand DNA copy of the element that is directly linked to the target DNA. In subsequent reactions, further processing of the single-strand DNA copy results in the generation of a new double-strand DNA copy of the *L1* element that is inserted at the site of the initial nick.

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 are retroviral-like retrotransposons. In *Drosophila*, DNA-based, retroviral, and nonretroviral 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 at Which 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 retrotransposons, 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 in 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 have 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 present, 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.

Conservative Site-Specific Recombination Can Reversibly Rearrange DNA

A different kind of recombination mechanism, known as *conservative site-specific recombination*, rearranges other types of mobile DNA elements. In this pathway, breakage and joining occur at two special sites, one on each participating DNA

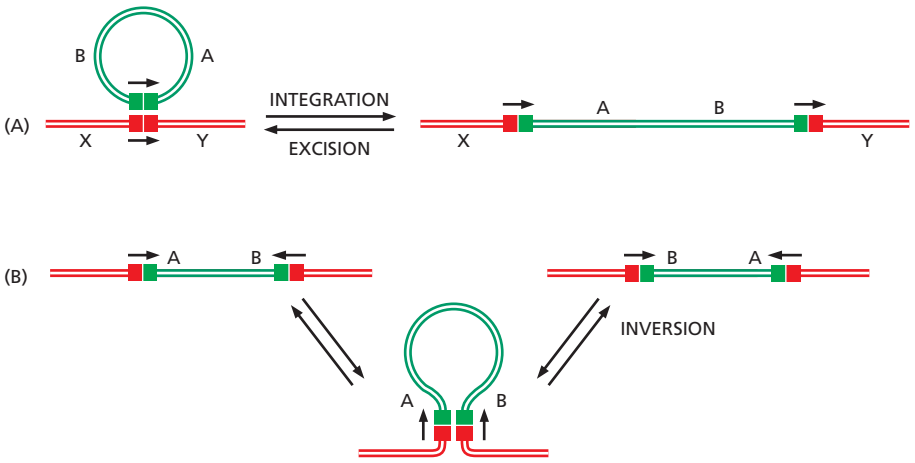


Figure 5–64 Two types of DNA rearrangement produced by conservative site-specific recombination. The only difference between the reactions in (A) and (B) is the relative orientation of the two short DNA sites (indicated by *arrows*) at which a site-specific recombination event occurs. (A) Through an integration reaction, a circular DNA molecule can become incorporated into a second DNA molecule; by the reverse reaction (excision), it can exit to re-form the original DNA circle. Many bacterial viruses move in and out of their host chromosomes in this way. (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; as described in the following section, the inversion of a DNA segment changes the type of flagellum that is produced by the bacterium.

molecule. Depending on the positions and relative orientations of the two recombination sites, DNA integration, DNA excision, or DNA inversion can occur (**Figure 5–64**). 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 5–64A).

Conservative site-specific recombination is often used by DNA viruses to move their genomes in and out of the genomes of their host cells. When integrated into its host genome, the viral DNA is replicated along with the host DNA and is faithfully passed on to all descendent cells. If the host cell suffers damage (for example, by UV irradiation), the virus can reverse the site-specific recombination reaction, excise its genome, and package it into a virus particle. In this way, many viruses can replicate themselves passively as a component of the host genome, but can also "leave the sinking ship" by excising their genomes and packaging them in a protective coat until a new, healthy host cell is encountered.

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 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 (see Figure 5–21). Thus, all the phosphate bonds that are broken during a recombination event are restored upon its completion (hence the term conservative). Transposition, in contrast, does not proceed through a covalently joined protein–DNA intermediate, and this process leaves gaps in the DNA that must be repaired by DNA polymerases.

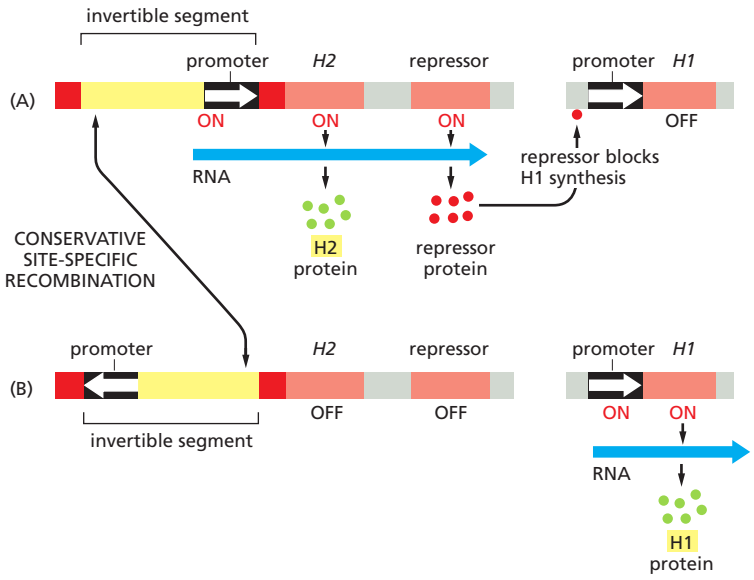


Figure 5–65 Switching gene expression by DNA inversion in bacteria. Alternating transcription of two flagellin genes in a *Salmonella* bacterium is caused by a conservative site-specific recombination event that inverts a small DNA segment containing a promoter. (A) In one orientation, the promoter activates transcription of the *H2* flagellin gene as well as that of a repressor protein that blocks the expression of the *H1* flagellin gene. Promoters and repressors are described in detail in Chapter 7; here we note simply that a promoter is needed to express a gene into protein and that a repressor blocks this from happening. (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 inversion reaction requires specific DNA sequences *(red)* and a recombinase enzyme that is encoded in the invertible DNA segment. This site-specific recombination mechanism is activated only rarely (about once in every 10^5 cell divisions). Therefore, the production of one or the other flagellin tends to be faithfully inherited in each clone of cells.

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

Many bacteria use conservative site-specific recombination to control the expression of particular genes. A well-studied example occurs in *Salmonella* bacteria and is known as **phase variation**. The switch in gene expression results from the occasional inversion of a specific 1000-nucleotide-pair piece of DNA, brought about by a conservative site-specific recombinase encoded in the *Salmonella* genome. This change alters the expression of the cell-surface protein flagellin, for which the bacterium has two different genes (**Figure 5–65**). The DNA inversion changes the orientation of a promoter (a DNA sequence that directs transcription of a gene) that is located within the inverted DNA segment. With the promoter in one orientation, the bacteria synthesize one type of flagellin; with the promoter in the other orientation, they synthesize the other type. The recombination reaction is reversible, allowing bacterial populations to switch back and forth between the two types of flagellin. Inversions occur only rarely, and because such changes in the genome will be copied faithfully during all subsequent replication cycles, entire clones of bacteria will have one type of flagellin or the other.

Phase variation helps protect 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.

Bacterial Conservative Site-Specific Recombinases Have Become Powerful Tools for Cell and Developmental Biologists

Like many of the mechanisms used by cells and viruses, site-specific recombination has been put to work by scientists to study a wide variety of problems. To decipher the roles of specific genes and proteins in complex multicellular organisms, genetic engineering techniques are used to produce worms, flies, and 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 a multicellular organism (**Figure 5–66**). 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

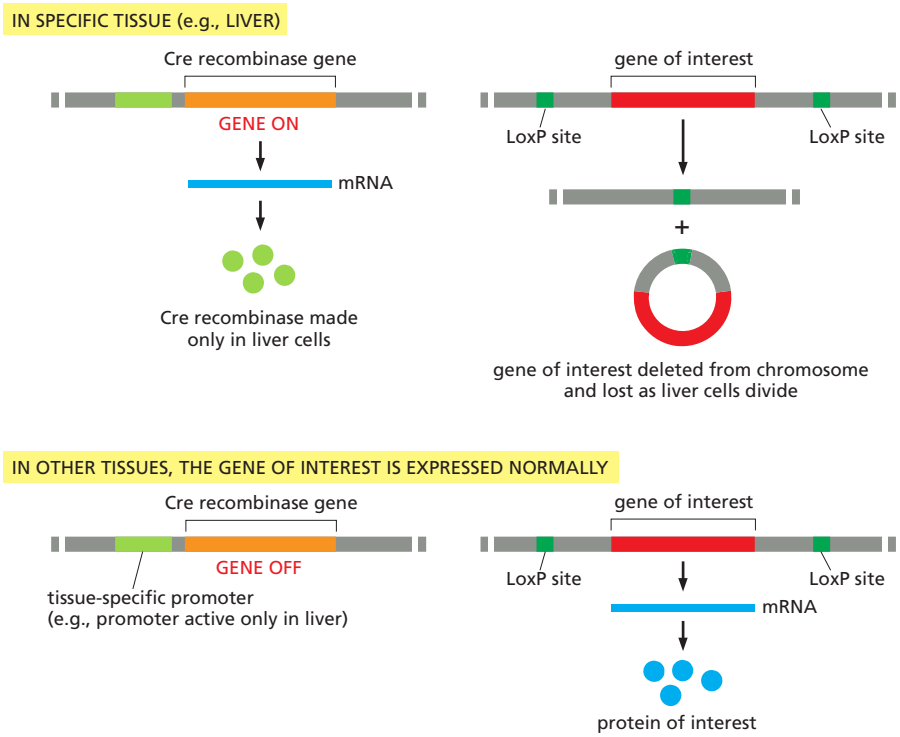


Figure 5–66 How a conservative site-specific recombination enzyme from bacteria is used to delete specific genes from particular mouse tissues. This approach requires the insertion of two specially engineered DNA molecules into the animal's germ line. The first contains the gene for a recombinase (in 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. The reaction that excises the gene is the same as that shown in Figure 5–64A. 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 different times during the development of each tissue.

very early in development. 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 played an important part in creating the genetic variation crucial for the evolution of cells and organisms.

WHAT WE DON'T KNOW

- How does DNA replication contend with all the other processes that occur simultaneously on chromosomes, including DNA repair and gene transcription?
- What is the basis for the low frequency of errors in DNA replication observed in all cells? Is this the best that cells can do given the speed of replication and the limits of molecular diffusion? Was this mutation rate selected in evolution to provide genetic variation?
- Cells have only one fundamental way of replicating DNA but many different ways of repairing it. Are there still other, undiscovered ways that cells have for repairing DNA?
- Do the many "dead" transposons in the human genome provide any benefits to humans?

PROBLEMS

Which statements are true? Explain why or why not.

5–1 The different cells in your body rarely have genomes with 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—in the absence of topoisomerase—would have to rotate at nearly 3000 revolutions per minute.

5–3 In a replication bubble, the same parental DNA strand serves as the template strand for leading-strand synthesis in one replication fork and as the template for lagging-strand synthesis in the other fork.

5–4 When bidirectional replication forks from adjacent origins meet, a leading strand always runs into a lagging strand.

5–5 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.

5–6 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 5–6)

	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

5–7 DNA repair enzymes preferentially repair mismatched bases on the newly synthesized DNA strand, using the old DNA strand as a template. If mismatches were instead repaired without regard for which strand served as template, would mismatch repair reduce replication errors? Would such a mismatch repair system 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.

5–8 Discuss the following statement: "Primase is a sloppy enzyme that makes many mistakes. Eventually, the RNA primers it makes are replaced with DNA made by a polymerase with higher fidelity. This is wasteful. It would be more energy-efficient if a DNA polymerase made an accurate copy in the first place."

5–9 If DNA polymerase requires a perfectly paired primer in order to add the next nucleotide, how is it that any mismatched nucleotides "escape" this requirement and become substrates for mismatch repair enzymes?

5–10 The laboratory you joined is studying the life cycle of an animal virus that uses circular, double-strand 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 broke open cells infected with the virus, isolated replicating viral genomes, cleaved them with a restriction nuclease that cuts the genome at only 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 relative to 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?

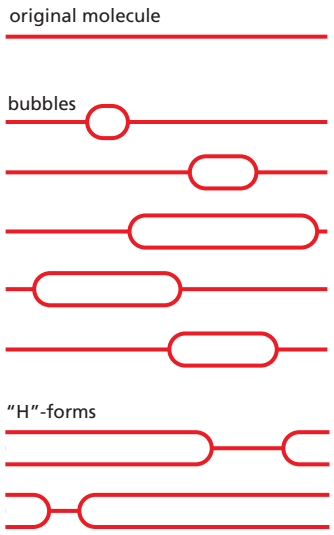


Figure Q5–1 Parental and replicating forms of an animal virus (Problem 5–10).

5–11 You are investigating DNA synthesis in tissue-culture cells, using ^{3}H-thymidine to radioactively label the replication forks. By breaking open the cells in a way that allows some of the DNA strands to be stretched out, very long DNA strands can be isolated intact and examined. You overlay the DNA with a photographic emulsion, and expose it for 3 to 6 months, a procedure known as autoradiography. Because the emulsion is sensitive to radioactive emissions, the ^{3}H-labeled DNA shows up as tracks of silver grains. Because the stretching collapses replication

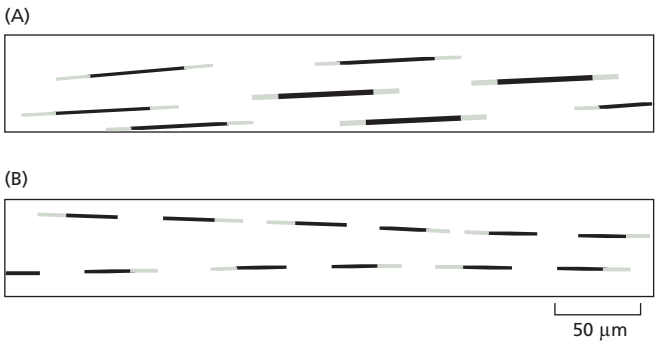


Figure Q5–2 Autoradiographic investigation of DNA replication in cultured cells (Problem 5–11). (A) Addition of ^{3}H-labeled thymidine immediately after release from the synchronizing block. (B) Addition of ^{3}H-labeled thymidine 30 minutes after release from the synchronizing block.

bubbles, the daughter duplexes lie side by side and cannot be distinguished from each other.

You pretreat the cells to synchronize them at the beginning of S phase. In the first experiment, you release the synchronizing block and add ^{3}H-thymidine immediately. After 30 minutes, you wash the cells and change the medium so that the total concentration of thymidine is the same as it was, but only one-third of it is radioactive. After an additional 15 minutes, you prepare DNA for autoradiography. The results of this experiment are shown in **Figure Q5–2A**. In the second experiment, you release the synchronizing block and then wait 30 minutes before adding ^{3}H-thymidine. After 30 minutes in the presence of ^{3}H-thymidine, you once again change the medium to reduce the concentration of radioactive thymidine and incubate the cells for an additional 15 minutes. The results of the second experiment are shown in **Figure Q5–2B**.

A. Explain why, in both experiments, some regions of the tracks are dense with silver grains (dark), whereas others are less dense (light).

B. In the first experiment, each track has a central dark section with light sections at each end. In the second experiment, the dark section of each track has a light section at only one end. Explain the reason for this difference.

C. Estimate the rate of fork movement (μm/min) in these experiments. Do the estimates from the two experiments agree? Can you use this information to gauge how long it would take to replicate the entire genome?

5–12 If you compare the frequency of the sixteen possible dinucleotide sequences in the *E. coli* and human genomes, 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?

5–13 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 protein-coding 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% protein-coding and 0.5% regulatory—is crucial information.)

5–14 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 Q5–3**. 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.

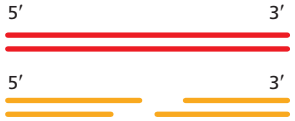


Figure Q5–3 A broken duplex with single-strand tails ready to invade an intact homologous duplex (Problem 5–14).

5–15 In addition to correcting DNA mismatches, the mismatch repair system functions to prevent homologous recombination from taking place between similar but not identical sequences. Why would recombination between similar, but nonidentical sequences pose a problem for human cells?

5–16 Cre recombinase is a site-specific enzyme that catalyzes recombination between two LoxP DNA sites. Cre recombinase pairs two LoxP sites in the same orientation, breaks both duplexes at the same point in each LoxP site, and joins the ends with new partners so that each LoxP site is regenerated, as shown schematically in **Figure Q5–4A**. Based on this mechanism, predict the arrangement of sequences that will be generated by Cre-mediated site-specific recombination for each of the two DNAs shown in **Figure Q5–4B**.

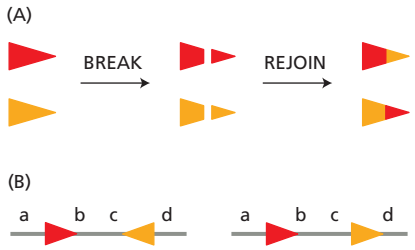


Figure Q5–4 Cre recombinase-mediated site-specific recombination (Problem 5–16). (A) Schematic representation of Cre/LoxP site-specific recombination. The LoxP sequences in the DNA are represented by *triangles* that are colored so that the site-specific recombination event can be followed more readily. In reality their DNA sequences are identical. (B) DNA substrates containing two arrangements of LoxP sites.

REFERENCES

General

Brown TA (2007) Genomes 3. New York: Garland Science.

Friedberg EC, Walker GC, Siede W et al. (2005) DNA Repair and Mutagenesis. Washington, DC: ASM Press.

Haber JE (2013) Genome Stability: DNA Repair and Recombination. New York: Garland Science.

Hartwell L, Hood L, Goldberg ML et al. (2010) 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. (2013) Molecular Biology of the Gene, 7th ed. Menlo Park, CA: Benjamin Cummings.

The Maintenance of DNA Sequences

Conrad DF, Keebler J, DePristo M et al. (2011) Variation in genome-wide mutation rates within and between human families. *Nat. Genet.* 43, 712–714.

Catarina D & Eichler EE (2013) Properties and rates of germline mutations in humans. *Trends Genet.* 29, 575–584.

Cooper GM, Brudno M, Stone ES et al. (2004) Characterization of evolutionary rates and constraints in three mammalian genomes. *Genome Res.* 14, 539–548.

Hedges SB (2002) The origin and evolution of model organisms. *Nat. Rev. Genet.* 3, 838–849.

King MC & Wilson AC (1965) Evolution at two levels in humans and chimpanzees. *Science* 188, 107–116.

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.

Kelch BA, Makino DL, O'Donnell M et al. (2011) How a DNA polymerase clamp loader opens a sliding clamp. *Science* 334, 1675–1680.

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, 6973–6977.

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.

O'Donnell M, Langston L & Stillman B (2013) Principals and concepts of DNA replication in Bacteria, Archaea, and Eukarya. *Cold Spring Harb. Lab. Perspect. Biol.* 195, 1231–1240.

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.

Vos SM, Tretter EM, Schmidt BH et al. (2011) All tangled up: how cells direct, manage and exploit topoisomerase function. *Nat. Rev. Mol. Cell Biol.* 12, 827–841.

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.

Gilbert DM (2010) Evaluating genome-scale approaches to eukaryotic DNA replication. *Nat. Rev. Genet.* 11, 673–684.

deLang T (2009) How telomeres solve the end-protection problem. *Science* 326, 948–952.

Mechali M (2010) Eukaryotic DNA replication origins: many choices for appropriate answers. *Nat. Rev. Mol. Cell Biol.* 11, 728–738.

Nandakumar J & Cech T (2013) Finding the end: recruitment of telomerase to telomeres. *Nat. Rev. Mol. Cell Biol.* 14, 69–82.

DNA Repair

Goodman MF & Woodgate, R (2013) Translesion DNA polymerases. *Cold Spring Harb. Perspect. Biol.* 5, a010363.

Hanawalt PC & Spivak G (2008) Transcription-coupled DNA repair: two decades of progress and surprises. *Nat. Rev. Mol. Cell Biol.* 9, 958–970.

Lindahl T (1993) Instability and decay of the primary structure of DNA. *Nature* 362, 709–715.

Malkova A & Haber JE (2012) Mutations arising during repair of chromosome breaks. *Annu. Rev. Genet.* 46, 455–473.

Prakash S, Johnson RE & Prakash L (2005) Eukaryotic translesion synthesis DNA polymerases: specificity of structure and function. *Annu. Rev. Biochem.* 74, 317–353.

Reardon JT & Sancar A (2005) Nucleotide excision repair. *Prog. Nucleic Acid Res. Mol. Biol.* 79, 183–235.

Homologous Recombination

Chen Z, Yang H & Pavletich NP (2008) Mechanism of homologous recombination from the RecA-ssDNA/dsDNA structures. *Nature* 453, 489–494.

Cox MM (2001) Historical overview: searching for replication help in all of the rec places. *Proc. Natl. Acad. Sci. USA* 98, 8173–8180.

Heyer WD, Ehmsen KT & Liu J (2010) Regulation of homologous recombination in eukaryotes. *Annu. Rev. Genet.* 44, 113–139.

Holliday R (1990) The history of the DNA heteroduplex. *BioEssays* 12, 133–142.

Hunter N (2006) Meiotic recombination. In *Topics in Current Genetics, Molecular Genetics of Recombination*, Aguilera A & Rothstein R (eds), pp. 381–422. Springer-Verlag: Heidelberg.

de Massy B (2013) Initiation of meiotic recombination: how and where? Conservation and specificities among eukaryotes. *Annu. Rev. Genet.* 47, 563–599.

Michel B, Gromponee G, Florès MJ & Bidnenko V (2004) Multiple pathways process stalled replication forks. *Proc. Natl. Acad. Sci. USA* 101, 12783–12788.

Moynahan ME & Jasin M (2010) Mitotic homologous recombination maintains genomic stability and suppresses tumorigenesis. *Nat. Rev. Mol. Cell Biol.* 11, 196–207.

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. *Nat. Rev. Mol. Cell Biol.* 4(6), 435–445.

Yeeles JY, Poli J, Marians KJ et al. (2013) Rescuing stalled or damaged replication forks. *Cold Spring Harb. Perspect. Biol.* 5, a012815.

Zickler D & Kleckner N (1999) Meiotic chromosomes: integrating structure and function. *Annu. Rev. Genet.* 33, 603–754.

Transposition and Conservative Site-specific Recombination

Comfort NC (2001) From controlling elements to transposons: Barbara McClintock and the Nobel Prize. *Trends Biochem. Sci.* 26, 454–457.

Grindley ND, Whiteson KL & Rice PA (2006) Mechanisms of site-specific recombination. *Annu. Rev. Biochem.* 75, 567–605.

Huang, CR, Burns KH & Boeke JD (2012) Active transposition in genomes. *Annu. Rev. Genet.* 46, 651–675.

Varmus H (1988) Retroviruses. *Science* 240, 1427–1435.

CHAPTER 6

How Cells Read the Genome: From DNA to Protein

IN THIS CHAPTER

Since the structure of DNA was discovered in the early 1950s, progress in cell and molecular biology has been astounding. We now know the complete genome sequences for thousands of different organisms, revealing fascinating details of their biochemistry as well as important clues as to how these organisms evolved. Complete genome sequences have also been obtained for thousands of individual humans, as well as for a few of our now-extinct relatives, such as the Neanderthals. Knowing the maximum amount of information that is required to produce a complex organism like ourselves puts constraints on the biochemical and structural features of cells and makes it clear that biology is not infinitely complex.

As discussed in Chapter 1, 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 6–1**). 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 of molecular biology, there are important variations between organisms in the way in which information flows from DNA to protein. Principal among these is that RNA transcripts in eukaryotic 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. As we discuss in this chapter, these processing steps can critically change the "meaning" of an RNA molecule and are therefore crucial for understanding how eukaryotic cells read their genome.

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, some of these RNAs fold into precise three-dimensional structures that have structural and catalytic roles in the cell. Other RNAs, as we discuss in the next chapter, act primarily as regulators of gene expression. But the roles of many noncoding RNAs are not yet known.

One might have predicted that the information present in genomes would be arranged in an orderly fashion, resembling a dictionary or a telephone directory. But it turns out that the genomes of most multicellular organisms are surprisingly disorderly, reflecting their chaotic evolutionary histories. The genes in these organisms largely consist of a long string of alternating short exons and long introns, as discussed in Chapter 4 (see Figure 4–15D). Moreover, small bits of DNA sequence that code 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 difficult for researchers to locate definitively the beginning and end of genes, much less to decipher when and where each gene is expressed in the life of the

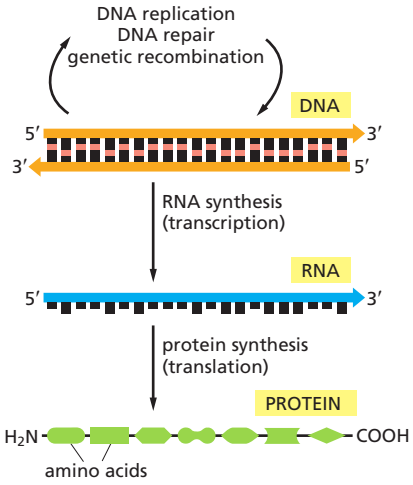


Figure 6–1 The pathway from DNA to protein. The flow of genetic information from DNA to RNA (transcription) and from RNA to protein (translation) occurs in all living cells.

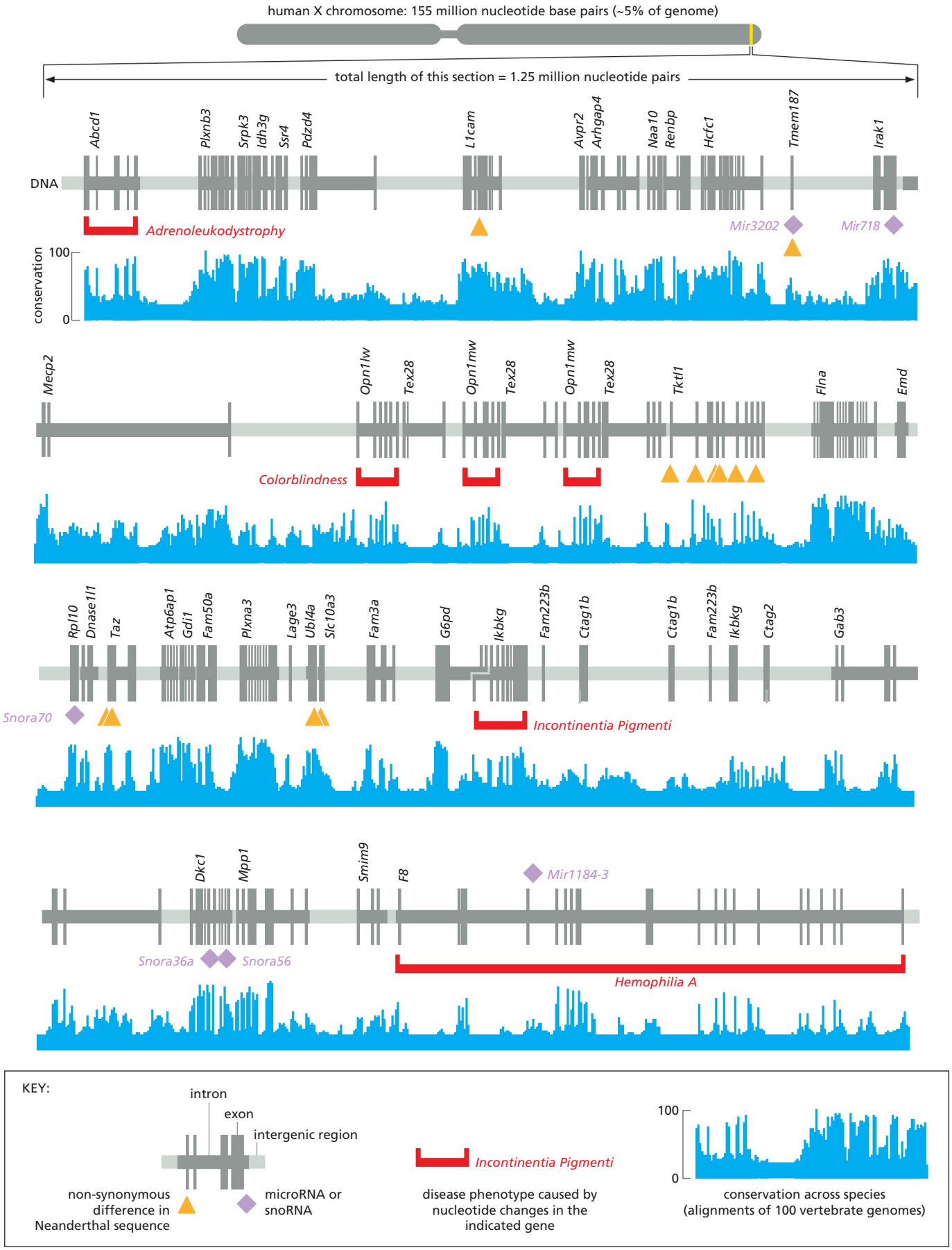

human X chromosome: 155 million nucleotide base pairs (~5% of genome)
total length of this section = 1.25 million nucleotide pairs
DNA
Abcd1
Plxnb3
Srpk3
Idh3g
Ssr4
Pdzd4
L1cam
Avpr2
Arhgap4
Naa10
Renbp
Hcfc1
Tmem187
Irak1
Adrenoleukodystrophy
Mir3202
Mir718
conservation
100
0
Mecp2
Opn1lw
Tex28
Opn1mw
Tex28
Opn1mw
Tex28
Tktl1
Flna
Emd
Colorblindness
Rpl10
Dnase1l1
Taz
Atp6ap1
Gdi1
Fam50a
Plxna3
Lage3
Ubl4a
Slc10a3
Fam3a
G6pd
Ikbkg
Fam223b
Ctag1b
Ctag1b
Fam223b
Ikbkg
Ctag2
Gab3
Snora70
Incontinentia Pigmenti
Dkc1
Mpp1
Smim9
F8
Mir1184-3
Snora36a
Snora56
Hemophilia A
KEY:
intron
exon
intergenic region
non-synonymous difference in Neanderthal sequence
microRNA or snoRNA
Incontinentia Pigmenti
disease phenotype caused by nucleotide changes in the indicated gene
100
0
conservation across species (alignments of 100 vertebrate genomes)

Figure 6–2 Schematic depiction of a small portion of the human X chromosome. As summarized in the key, the known protein-coding genes (starting with *Abcd1* and ending with *F8*) are shown in *dark gray*, with coding regions (exons) indicated by bars that extend above and below the central line. Noncoding RNAs with known functions are indicated by *purple diamonds*. *Yellow triangles* indicate positions within protein-coding regions where the Neanderthal genome sequences codes for a different amino acid than the human genome. The stretch of *yellow triangles* in the *Txtl1* gene appear to have been positively selected for since the divergence of *Homo sapiens* from Neanderthals some 200,000 years ago. Note that most of the proteins are identical between us and our extinct relative. The *blue* histogram indicates the extent to which portions of the human genome are conserved with other vertebrate species. It is likely that additional genes, currently unrecognized, also lie within this portion of the human genome.

Genes whose mutation causes an inherited human condition are indicated by *red brackets*. The *Abcd1* gene codes for a protein that imports fatty acids into the peroxisome; mutations in the gene cause demylination of nerves which can result in cognition and movement disorders. *Incontinentia pigmenti* is a disease of the skin, hair, nails, teeth, and eyes. *Hemophilia A* is a bleeding disorder caused by mutations in the Factor VIII gene, which codes for a blood-clotting protein. Because males have only a single copy of the X chromosome, most of the conditions shown here affect only males; females that inherit one of these defective genes are often asymptomatic because a functional protein is made from their other X chromosome. (Courtesy of Alex Williams, obtained from the University of California, Genome Browser, http://genome.ucsc.edu)

organism. Yet the cells in our body do this automatically, thousands of times a second.

The problems that cells face in decoding genomes can be appreciated by considering a tiny portion of the human genome (**Figure 6–2**). The region illustrated represents less than 1/2000th of our genome and includes at least 48 genes that encode proteins and 6 genes for noncoding RNAs. When we consider the entire human genome, we can only marvel at the capacity of our cells to rapidly and accurately handle such large amounts of information.

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 30,000 genes. An enormous amount of ignorance remains; many fascinating challenges therefore await the next generation of cell biologists.

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 from a gene when necessary. But genes can be transcribed and translated with different efficiencies, allowing the cell to make vast quantities of some proteins and tiny amounts of others (**Figure 6–3**). Moreover, as we see in the next chapter,

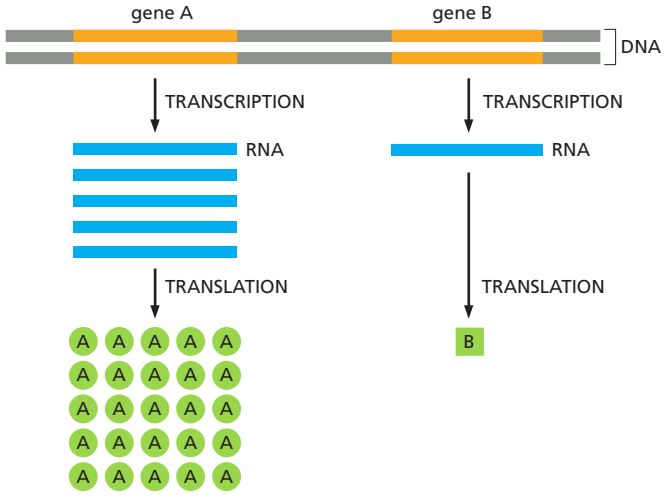


Figure 6–3 Genes can be expressed with different efficiencies. In this example, gene A is transcribed much more efficiently than gene B and each RNA molecule that it produces is also translated more frequently. This causes the amount of protein A in the cell to be much greater than that of protein B.

a cell can change (or regulate) the expression of each of its genes according to its needs—most commonly by controlling the production of its RNA.

RNA Molecules Are Single-Stranded

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 (**Figure 6–4**). 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 given to producing RNA molecules on DNA is *transcription*.

Like DNA, RNA is a linear polymer made of four different types of nucleotide subunits linked together by phosphodiester bonds (see 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 (**Figure 6–5**). Since U, like T, can base-pair by hydrogen-bonding with A (**Figure 6–6**), 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.

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 6–7**). 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.

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

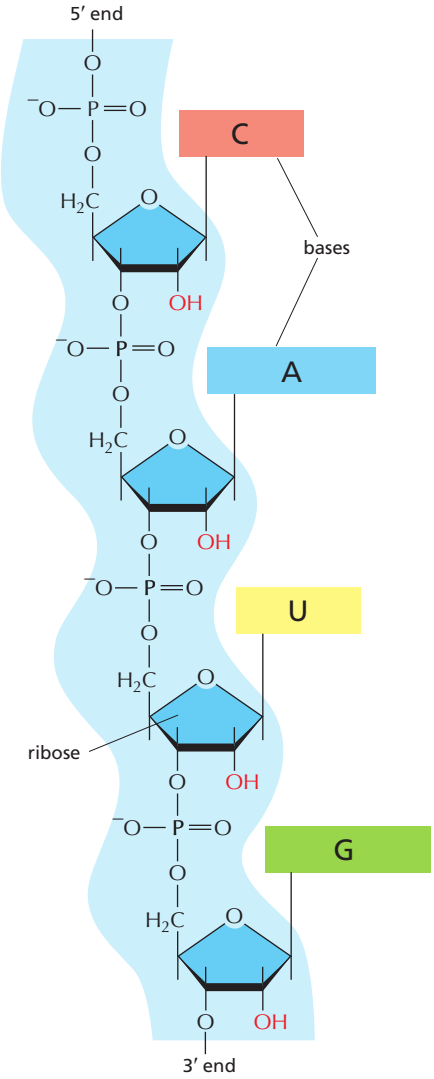


Figure 6–4 A short length of RNA. The phosphodiester chemical linkage between nucleotides in RNA is the same as that in DNA.

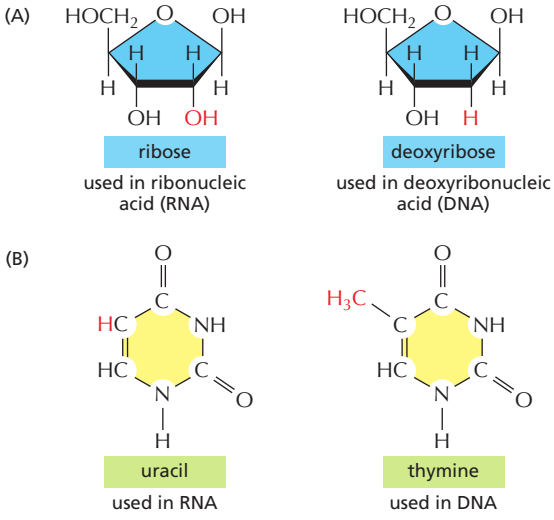


Figure 6–5 The chemical structure of RNA. (A) RNA contains the sugar ribose, which differs from deoxyribose, the sugar used in DNA, by the presence of an additional –OH group. (B) RNA contains the base uracil, which differs from thymine, the equivalent base in DNA, by the absence of a $–CH_3$ group.

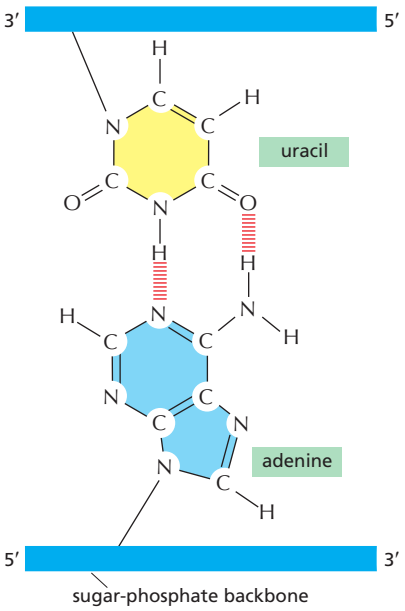


Figure 6–6 Uracil forms base pairs with adenine. The absence of a methyl group in U has no effect on base-pairing; thus, U-A base pairs closely resemble T-A base pairs (see Figure 4–4).

template. When a good match is made (A with T, U with A, G with C, and C with G), 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 6–8**).

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.

RNA Polymerases Carry Out Transcription

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

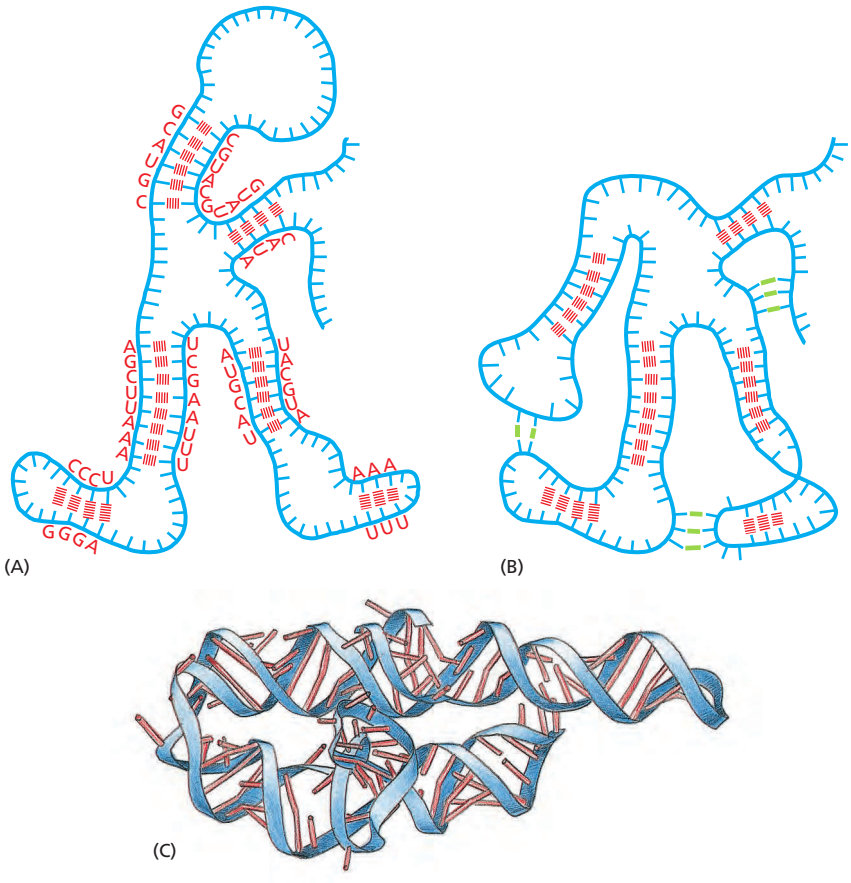


Figure 6–7 RNA can fold into specific structures. RNA is largely single-stranded, but it often contains short stretches of nucleotides that can form conventional base pairs with complementary sequences found elsewhere on the same molecule. These interactions, along with additional "nonconventional" base-pair interactions, allow an RNA molecule to fold into a three-dimensional structure that is determined by its sequence of nucleotides (**Movie 6.1**). (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, one that catalyzes its own splicing (see p. 324). Each conventional base-pair interaction is indicated by a "rung" in the double helix. Bases in other configurations are indicated by broken rungs.

base-pairing. In this way, the growing RNA chain is extended by one nucleotide at a time in the 5′-to-3′ direction (Figure 6–9). The substrates are ribonucleoside 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 and Movie 6.2).

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 previous RNA molecules are completed (Figure 6–10). When RNA polymerase molecules follow hard on each other's heels in this way, each moving at about 50 nucleotides per second, 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 is thought possible because transcription need not be as accurate as DNA replication (see Table 5–1, p. 244). 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 as RNA does not permanently store genetic information in cells. Finally, unlike DNA polymerases, which make their products in segments that are later stitched together, RNA polymerases are absolutely processive; that is, the same RNA polymerase that begins an RNA molecule must finish it without dissociating from the DNA template.

Although not nearly as accurate as the DNA polymerases that replicate DNA, RNA polymerases 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, except that a water molecule replaces the pyrophosphate 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 reveal that, other than containing a critical Mg^{2+} ion at the catalytic site, the two enzymes are quite different. Template-dependent nucleotide-polymerizing enzymes seem to have arisen at least twice during the early evolution of cells. One lineage led to the

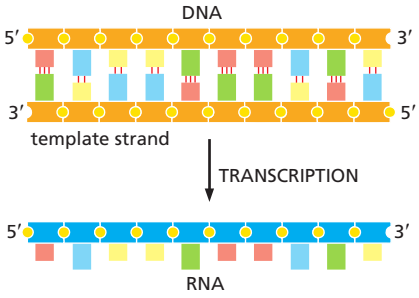


Figure 6–8 DNA transcription produces a single-stranded RNA molecule that is complementary to one strand of the DNA double helix. Note that the sequence of bases in the RNA molecule produced is the same as the sequence of bases in the non-template DNA strand, except that a U replaces every T base in the DNA.

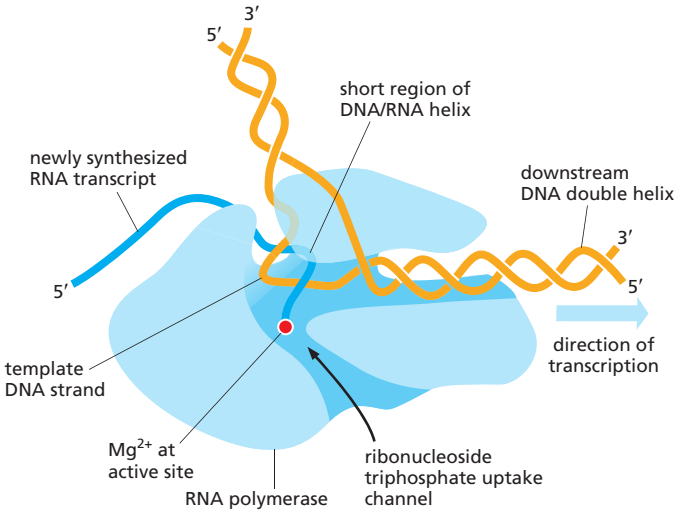


Figure 6–9 DNA is transcribed by the enzyme RNA polymerase. The RNA polymerase *(pale blue)* moves stepwise along the DNA, unwinding the DNA helix at its active site indicated by the Mg^{2+} *(red)*, which is required for catalysis. As it progresses, the polymerase adds nucleotides one by one to the 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 formed only transiently, and a "window" of DNA/RNA helix therefore moves along the DNA with the polymerase as the DNA double helix reforms behind it. 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). The figure, based on an x-ray crystallographic structure, shows a cut-away view of the polymerase: the part facing the viewer has been sliced away to reveal the interior (Movie 6.3). (Adapted from P. Cramer et al., *Science* 288:640–649, 2000; PDB code: 1HQM.)

Figure 6–10 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 relative 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.)

modern DNA polymerases and reverse transcriptases discussed in Chapter 5, as well as to a few RNA polymerases from viruses. The other lineage formed all of the modern RNA polymerases that we discuss in this chapter.

Cells Produce Different Categories of RNA Molecules

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 other genes, however, is the RNA molecule itself. These RNAs are known as **noncoding RNAs** because they do not code for protein. In a well-studied, single-celled eukaryote, the yeast *Saccharomyces cerevisiae*, over 1200 genes (more than 15% of the total) produce RNA as their final product. Humans may produce on the order of ten thousand noncoding RNAs. These RNAs, like proteins, serve as enzymatic, structural, and regulatory components for a wide variety of processes in the cell. In Chapter 5, we encountered one of them as the template carried by the enzyme telomerase. Although many of the 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. In Chapter 7, we shall see that *microRNA* (*miRNA*) molecules and *small interfering RNA* (*siRNA*) molecules serve as key regulators of eukaryotic gene expression, and that *piwi-interacting RNAs* (*piRNAs*) protect animal germ lines from transposons; we also discuss the *long noncoding RNAs* (*lncRNAs*), a diverse set of RNAs whose functions are just being discovered (**Table 6–1**).

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, help to process and chemically modify rRNAs
miRNAs	MicroRNAs, regulate gene expression by blocking translation of specific mRNAs and cause their degradation
siRNAs	Small interfering RNAs, turn off gene expression by directing the degradation of selective mRNAs and the establishment of compact chromatin structures
piRNAs	Piwi-interacting RNAs, bind to piwi proteins and protect the germ line from transposable elements
lncRNAs	Long noncoding RNAs, many of which serve as scaffolds; they regulate diverse cell processes, including X-chromosome inactivation

Each transcribed segment of DNA is called a *transcription unit*. In eukaryotes, 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, whereas proteins comprise about 50%. 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 eukaryotes. 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 synthesizes RNA using the DNA template as a guide. An additional 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 the two collide, and a holoenzyme typically slides rapidly along the long DNA molecule and then dissociates. However, when the polymerase holoenzyme slides into a special sequence of nucleotides indicating the starting point for RNA synthesis called a **promoter**, the polymerase binds tightly, because its σ factor makes specific contacts with the edges of bases exposed on the outside of the DNA double helix (step 1 in Figure 6–11A).

The tightly bound RNA polymerase holoenzyme at a promoter opens up the double helix to expose a short stretch of nucleotides on each strand (step 2 in Figure 6–11A). The region of unpaired DNA (about 10 nucleotides) is called the *transcription bubble* and it is stabilized by the binding of σ factor to the unpaired bases on one of the exposed strands. The other exposed DNA strand then 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 6–11A). The first ten or so nucleotides of RNA are synthesized using a "scrunching" mechanism, in which RNA polymerase remains bound to the promoter and pulls the upstream DNA into its active site, thereby expanding the transcription bubble. This process creates considerable stress and the short RNAs are often released, thereby relieving the stress and forcing the polymerase, which remains in place, to begin synthesis over again. Eventually this process of *abortive initiation* is overcome and the stress generated by scrunching helps the core enzyme to break free of its interactions with the promoter DNA (step 4 in Figure 6–11A) and discard the σ factor (step 5 in Figure 6–11A). At this point, the polymerase begins to move down the DNA, synthesizing RNA, in a stepwise fashion: the polymerase moves forward one base pair for every nucleotide added. During this process, the transcription bubble continually expands at the front of the polymerase and contracts at its rear. Chain elongation continues (at a speed of approximately 50 nucleotides/sec for bacterial RNA polymerases) until the enzyme encounters a second signal, the **terminator** (step 6 in Figure 6–11A), where the polymerase halts and releases both the newly made RNA molecule and the DNA template (step 7 in Figure 6–11A). The free polymerase core enzyme then reassociates with a free σ factor to form a holoenzyme that can begin the process of transcription again (step 8 in Figure 6–11A).

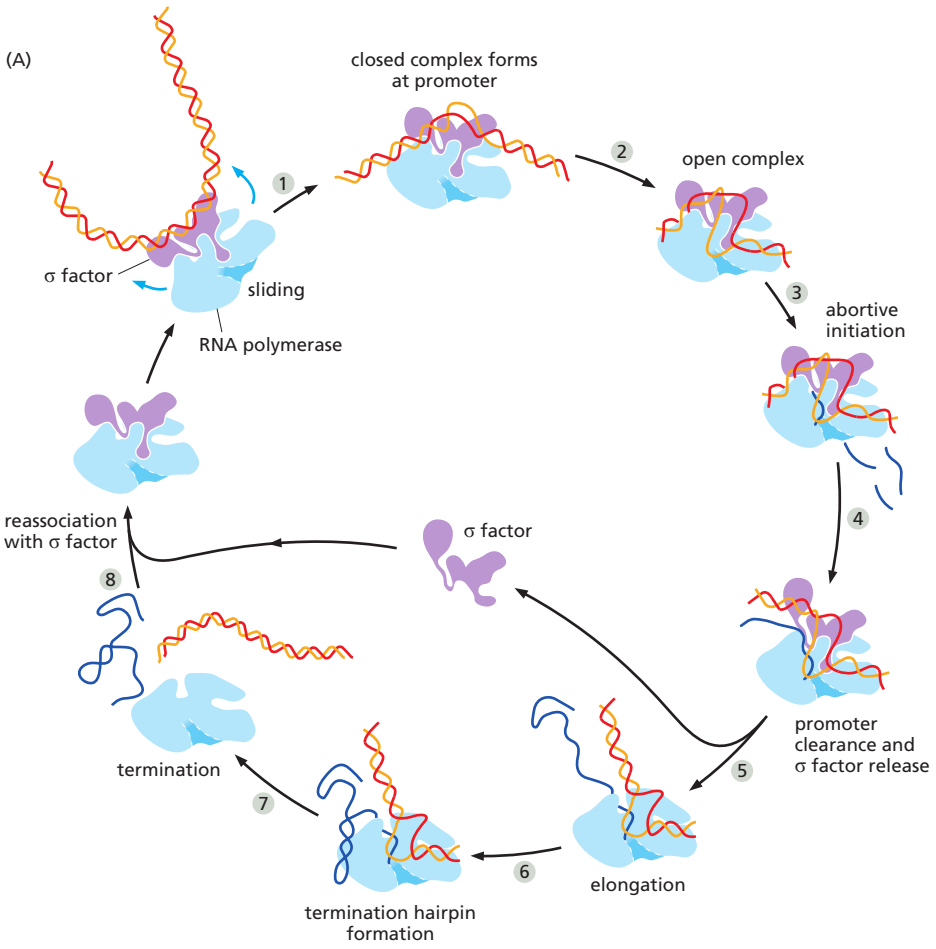


Figure 6–11 The transcription cycle of bacterial RNA polymerase. (A) In step 1, the RNA polymerase holoenzyme (polymerase core enzyme plus σ factor) assembles and then locates a promoter DNA sequence (see Figure 6–12). The polymerase opens (unwinds) the DNA at the position at which transcription is to begin (step 2) and begins transcribing (step 3). This initial RNA synthesis (abortive initiation) is relatively inefficient as short, unproductive transcripts are often released. However, once RNA polymerase has managed to synthesize about 10 nucleotides of RNA, it breaks its interactions with the promoter DNA (step 4) and eventually releases σ factor—as the polymerase tightens around the DNA and shifts to the elongation mode of RNA synthesis, moving along the DNA (step 5). During the elongation mode, 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 hairpin-like structure that destabilizes the polymerase's hold on the RNA.

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. (B) Two-dimensional image of an elongating bacterial RNA polymerase, as determined by atomic force microscopy (see Figure 9–33). (C) Interpretation of the image in (B). (Adapted from K.M. Herbert et al., *Annu. Rev. Biochem.* 77:149–176, 2008.)

The process of transcription initiation is complicated 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 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 twofold symmetric DNA sequence, which, when transcribed into RNA, folds into a "hairpin" structure through Watson–Crick base-pairing (see Figure 6–11A). As the polymerase transcribes across a terminator, the formation of the hairpin helps to disengage the RNA transcript from the active site (step 7 in Figure 6–11A). The process of termination provides 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 aspects of the DNA that are recognized directly

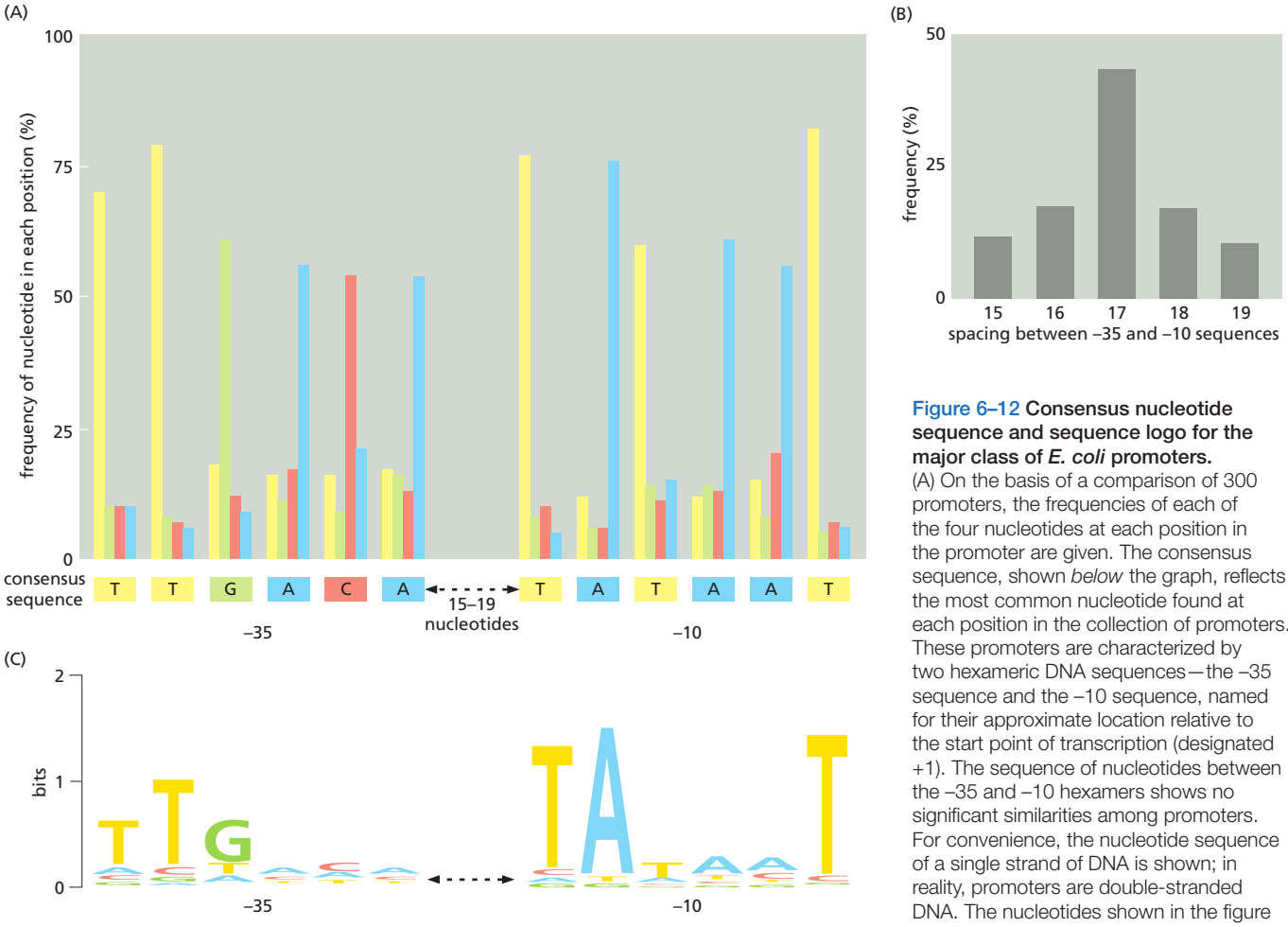


Figure 6–12 Consensus nucleotide sequence and sequence logo for the major class of *E. coli* promoters. (A) On the basis of a comparison of 300 promoters, the frequencies of each of the four nucleotides at each position in the promoter are given. The consensus sequence, shown *below* the graph, reflects the most common nucleotide found at each position in the collection of promoters. These 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). The sequence of nucleotides between the –35 and –10 hexamers shows no significant similarities among promoters. For convenience, the nucleotide sequence of a single strand of DNA is shown; in reality, promoters are double-stranded DNA. The nucleotides shown in the figure are recognized by σ factor, a subunit of the RNA polymerase holoenzyme. (B) The distribution of spacing between the –35 and –10 hexamers found in *E. coli* promoters. (C) A *sequence logo* displaying the same information as in panel (A). Here, the height of each letter is proportional to the frequency at which that base occurs at that position across a wide variety of promoter sequences. The total height of all the letters at each position is proportional to the information content (expressed in bits) at that position. For example, the total information content of a position that can tolerate several different bases is small (see the last three bases of the –35 sequences), but statistically greater than random.

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 nucleotides found at each position. It therefore serves as a summary or "average" of a large number of individual nucleotide sequences. A more accurate way of displaying the range of DNA sequences recognized by a protein is through the use of a *sequence logo*, which reveals the relative frequencies of each nucleotide at each position (Figure 6–12C).

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 promoter strengths. Promoters for genes that code for abundant proteins are much stronger than those associated with genes that encode rare proteins, and the nucleotide sequences of their promoters 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

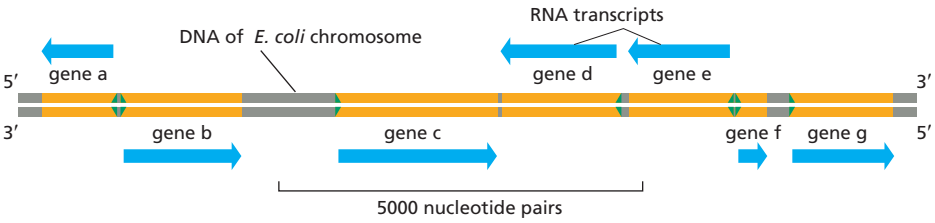


Figure 6–13 Directions of transcription along a short portion of a bacterial chromosome. Some genes are transcribed using one DNA strand as a template, while others are transcribed using the other DNA strand. The direction of transcription is determined by the promoter at the beginning of each gene *(green 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.

by analysis of the nucleotide sequence of a genome. It is even more difficult to locate analogous sequences in eukaryotic genomes, due in part to the excess DNA carried in these genomes. Often we need additional information, some of it from direct experimentation, to locate and accurately interpret the short DNA signals in genomes.

As shown in Figure 6–11, promoter sequences are asymmetric, ensuring that RNA polymerase can bind in only one orientation. Because the polymerase can synthesize RNA only in the 5′-to-3′ direction, the promoter orientation specifies the strand to be used as a template. Genome sequences reveal that the DNA strand that is used as the template for RNA synthesis varies from gene to gene, depending on the orientation of the promoter (**Figure 6–13**).

Having considered transcription in bacteria, we now turn to the situation in eukaryotes, where the synthesis of RNA molecules is a much more elaborate affair.

Transcription Initiation in Eukaryotes Requires Many Proteins

In contrast to bacteria, which contain a single type of RNA polymerase, eukaryotic nuclei have three: *RNA polymerase I*, *RNA polymerase II*, and *RNA polymerase III*. The three polymerases are structurally similar to one another and share some common subunits, but they transcribe different categories 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.

Eukaryotic RNA polymerase II has many structural similarities to bacterial RNA polymerase (**Figure 6–14**). But there are several important differences in the way in which the bacterial and eukaryotic enzymes function, two of which concern us immediately.

1. While bacterial RNA polymerase requires only a single transcription- initiation factor (σ) to begin transcription, eukaryotic RNA polymerases require many such factors, collectively called the *general transcription factors.*
2. Eukaryotic transcription initiation must take place on DNA that is packaged into nucleosomes and higher-order forms of chromatin structure (described in Chapter 4), features that are absent from bacterial chromosomes.

TABLE 6–2 The Three RNA Polymerases in Eukaryotic 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, lncRNA genes, and most snRNA genes
RNA polymerase III	tRNA genes, 5S rRNA genes, some snRNA genes, and genes for other small RNAs

The rRNAs were 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.

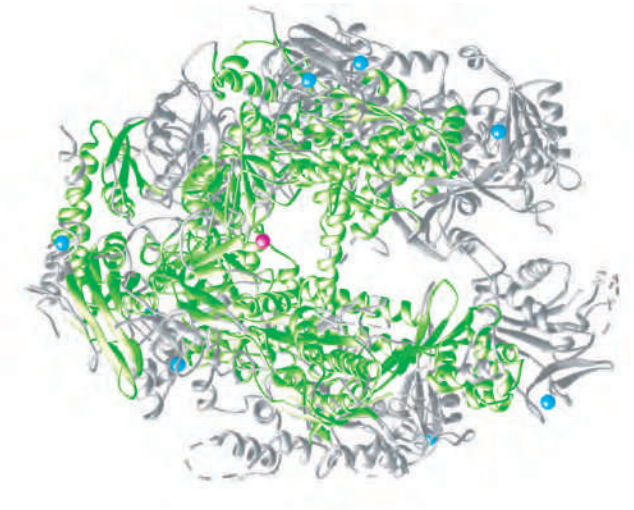


Figure 6–14 Structural similarity between a bacterial RNA polymerase and a eukaryotic RNA polymerase II. Regions of the two RNA polymerases that have similar structures are indicated in *green*. The eukaryotic polymerase is larger than the bacterial enzyme (12 subunits instead of 5), and some of the 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 eukaryotes) 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.)

RNA Polymerase II Requires a Set of General Transcription Factors

The **general transcription factors** help to position eukaryotic 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 to start its elongation mode. The proteins are "general" because they are needed at nearly all promoters used by RNA polymerase II. They consist of a set of interacting proteins denoted arbitrarily as TFIIA, TFIIB, TFIIC, TFIID, and so on (TFII standing for "transcription factor for polymerase II)." In a broad sense, the eukaryotic 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 6–15 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 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 6–16**), but for most polymerase II promoters it is the most important. The binding of TFIID

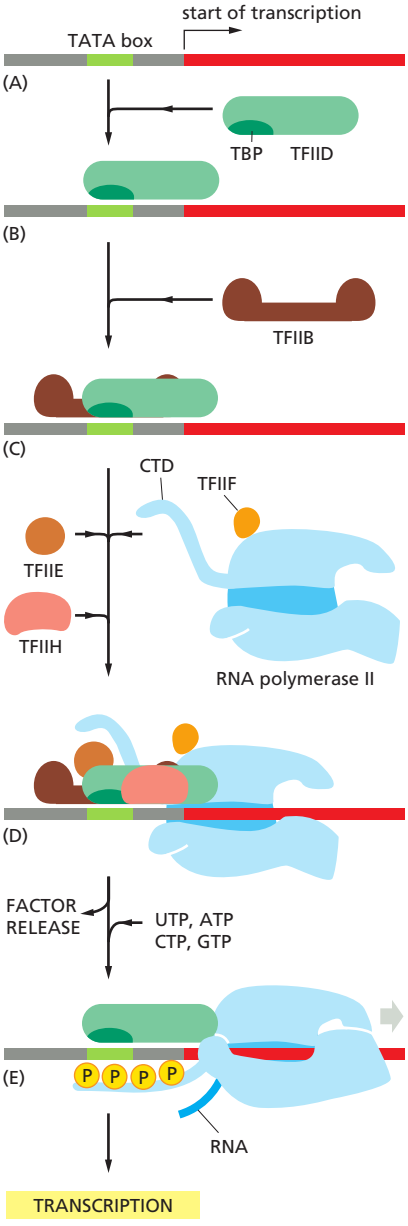


Figure 6–15 Initiation of transcription of a eukaryotic gene by RNA polymerase II. To begin transcription, RNA polymerase requires several general transcription factors. (A) The promoter contains a DNA sequence called the TATA box, which is located 25 nucleotides away from the site at which transcription 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 6–17) 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 energy from ATP hydrolysis 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 probably varies from gene to gene *in vivo*. The general transcription factors are highly conserved; some of those from human cells can be replaced in biochemical experiments by the corresponding factors from simple yeasts.

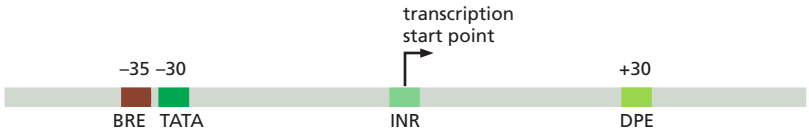
TABLE 6–3 The General Transcription Factors Needed for Transcription Initiation by Eukaryotic RNA Polymerase II

Name	Number of subunits	Roles in transition initiation
TFIID TBP subunit TAF subunits	 1 ~11	 Recognizes TATA box 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.

causes a large distortion in the DNA of the TATA box (Figure 6–17). 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 closer 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 6–15). The most complicated of the general transcription factors is TFIIH. Consisting of nine 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



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 subunit of TFIID
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 6–16 Consensus sequences found in the vicinity of eukaryotic 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.

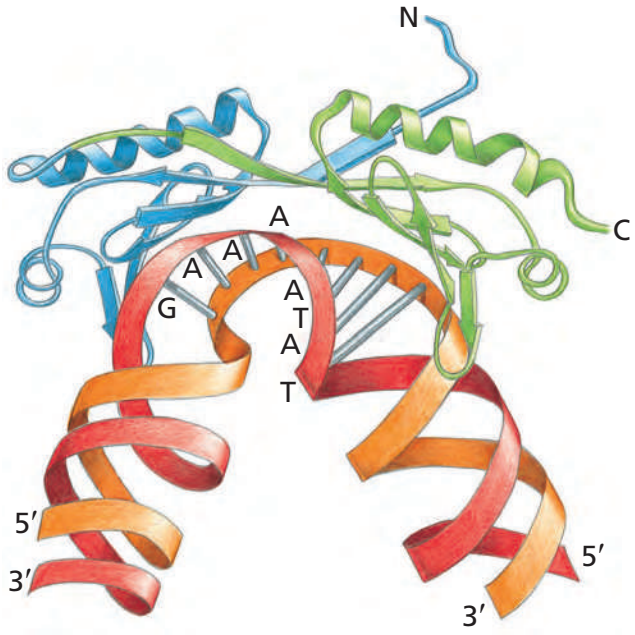


Figure 6–17 Three-dimensional structure of TBP (TATA-binding protein) bound to DNA. The TBP is the subunit of the general transcription factor TFIID that is responsible for recognizing and binding to the TATA box sequence in the DNA *(red).* The unique DNA bending caused by TBP—kinks in the double helix separated by partly unwound DNA—is thought to serve as a landmark that helps to attract the other general transcription factors (**Movie 6.4**). 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.)

seven-amino-acid sequence, which extend from the RNA polymerase core structure. During transcription initiation, the serine located at the fifth position in the repeat sequence (Ser5) is phosphorylated by TFIIH, which contains a protein kinase in one of its subunits (see Figure 6–15D 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, 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 has an additional function: it 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 RNA polymerase II and its general transcription factors acting on DNA templates in purified *in vitro* systems established the model for transcription initiation just described. However, as discussed in Chapter 4, DNA in eukaryotic cells is packaged into nucleosomes, which are further arranged in higher-order chromatin structures. As a result, transcription initiation in a eukaryotic cell is more complex and requires more proteins than it does on purified DNA. First, gene regulatory proteins known as *transcriptional activators* must bind to specific sequences in DNA (called *enhancers*) and help to attract RNA polymerase II to the start point of transcription (**Figure 6–18**). We discuss the role of these 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 eukaryotic cell. Second, eukaryotic transcription initiation *in vivo* requires the presence of a large 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 eukaryotic cell typically requires the recruitment of chromatin-modifying enzymes, including chromatin remodeling complexes and

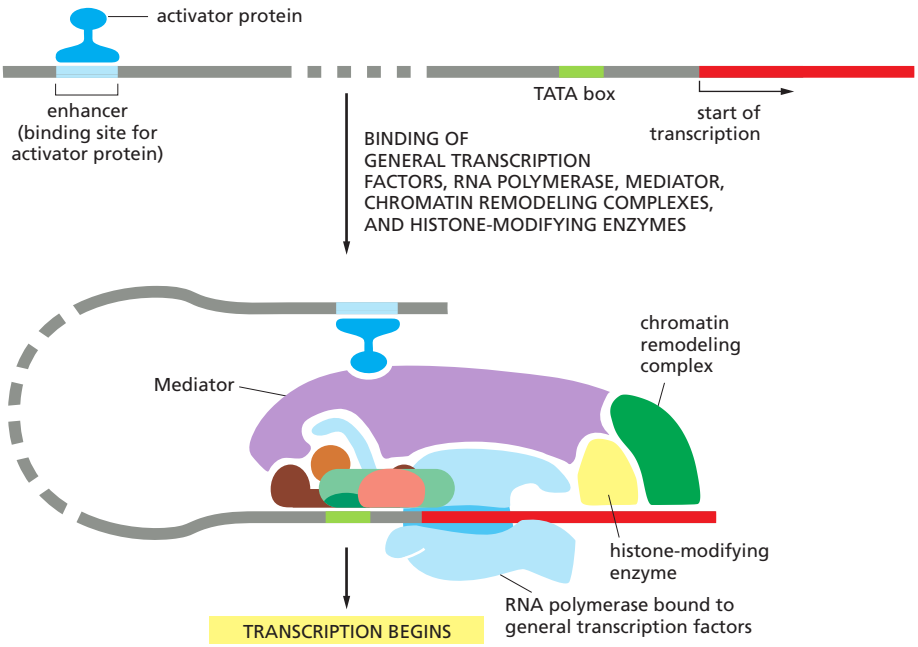


Figure 6–18 Transcription initiation by RNA polymerase II in a eukaryotic cell. Transcription initiation *in vivo* requires the presence of transcription activator proteins. As described in Chapter 7, these proteins bind to specific short sequences in DNA. Although only one is shown here, a typical eukaryotic gene utilizes many transcription activator proteins, which in combination determine its rate and pattern of transcription. Sometimes acting from a distance of several thousand nucleotide pairs (indicated by the dashed DNA molecule), these proteins help RNA polymerase, the general transcription factors, and Mediator all to assemble at the promoter. In addition, activators attract ATP-dependent chromatin remodeling complexes and histone-modifying enzymes. One of the main roles of Mediator is to coordinate the assembly of all these proteins at the promoter so that transcription can begin. As discussed in Chapter 4, the "default" state of chromatin is a condensed fiber (see Figure 4–28), and this is likely to be the form of DNA upon which most transcription is initiated. For simplicity, the chromatin is not shown in this figure.

histone-modifying enzymes. As discussed in Chapter 4, both types of enzymes can increase access to the DNA in chromatin, and by doing so they facilitate the assembly of the transcription initiation machinery onto DNA.

As illustrated in Figure 6–18, many proteins (well over 100 individual subunits) must assemble at the start point of transcription to initiate transcription in a eukaryotic 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 be brought to DNA as preformed subassemblies.

To begin transcribing, RNA polymerase II must be released from this large complex of proteins. In addition to the steps described in Figure 6–14, this release often requires the *in situ* proteolysis of the activator protein. We shall return to some of these issues, including the role of chromatin remodeling complexes and histone-modifying enzymes, in Chapter 7, where we discuss how eukaryotic cells regulate the process of transcription initiation.

Transcription Elongation in Eukaryotes Requires Accessory Proteins

Once RNA polymerase has initiated transcription, it moves jerkily, pausing at some DNA sequences and rapidly transcribing through others. Elongating RNA polymerases, both bacterial and eukaryotic, 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 shortly after initiation and help the polymerase move through the wide variety of different DNA sequences that are found in genes. Eukaryotic 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 that either move with the polymerase or may simply seek out and rescue the occasional stalled polymerase. In addition, histone chaperones help by partially disassembling nucleosomes in front of a moving RNA polymerase and assembling them behind.

As RNA polymerase moves along a gene, some of the enzymes bound to it modify the histones, leaving behind a record of where the polymerase has been. Although it is not clear exactly how the cell uses this information, it may aid in

transcribing a gene over and over again once it has become active for the first time. It may also be used to coordinate transcription elongation with the processing of RNA as it emerges from RNA polymerase, a topic we discuss later in this chapter.

Transcription Creates Superhelical Tension

There is yet another barrier to elongating RNA polymerases, both bacterial and eukaryotic, one that also applies to DNA polymerases, as discussed in Chapter 5 (see Figure 5–20). To describe this issue in more detail, we need first to consider a subtle property inherent in the DNA double helix called **DNA supercoiling**. DNA supercoiling is the name given to a conformation that DNA adopts in response to superhelical tension; alternatively, creating loops or coils in a double-helical DNA molecule can create such tension. **Figure 6–19** illustrates why. There are approximately 10 nucleotide pairs for every helical turn in a DNA double helix. If we 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 eukaryotic chromosomes), 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 creates superhelical tension as it moves along a stretch of DNA that is anchored at its ends (see Figure 6–19C). 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 positive superhelical tension in the DNA in front of it and negative helical tension behind it. For eukaryotes, this situation is thought to provide a bonus: although the positive superhelical tension ahead of the polymerase makes the DNA helix

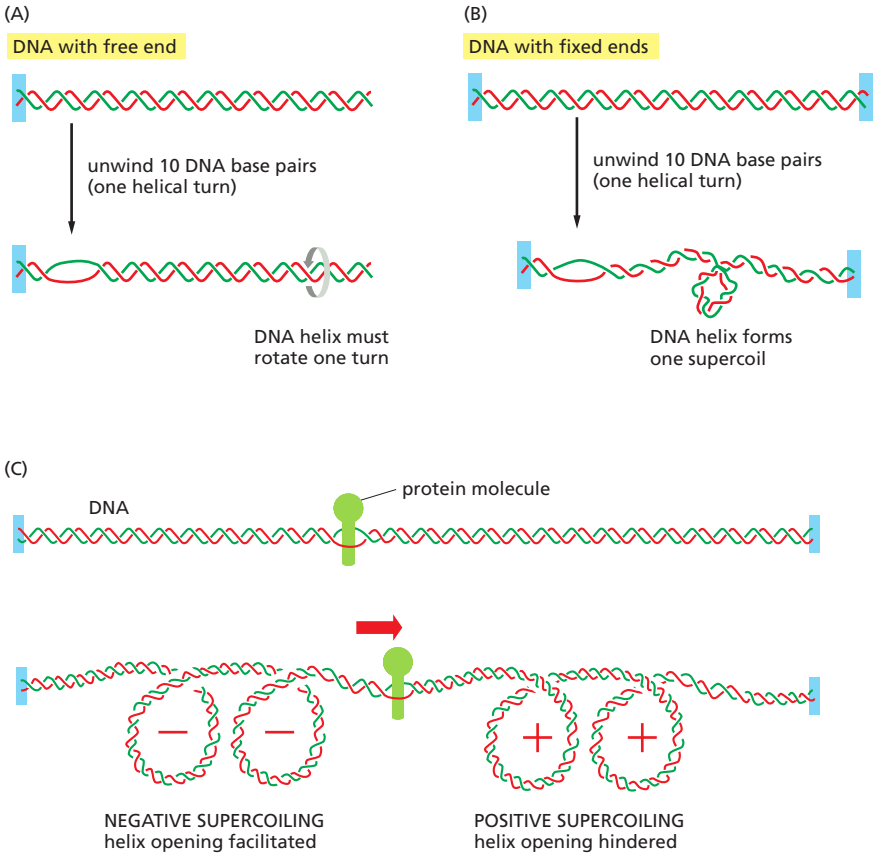


Figure 6–19 Superhelical tension in DNA causes DNA supercoiling. (A) For a DNA molecule with one free end (or a nick in one strand that serves as a swivel), the DNA double helix rotates by one turn for every 10 nucleotide pairs opened. (B) If rotation is prevented, superhelical tension is introduced into the DNA by helix opening. In the example shown, the DNA helix contains 10 helical turns, one of which is opened. One way of accommodating the tension created 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.

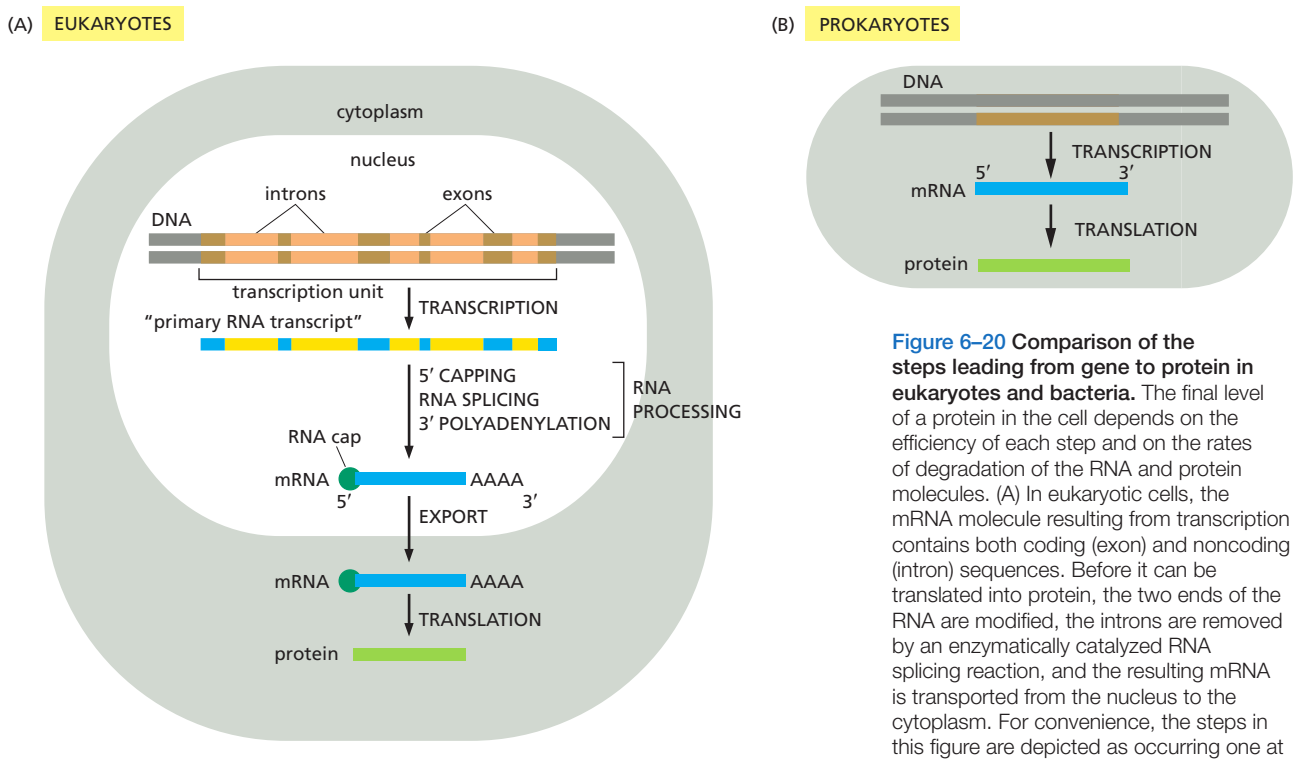


Figure 6–20 Comparison of the steps leading from gene to protein in eukaryotes and bacteria. The final level of a protein in the cell depends on the efficiency of each step and on the rates of degradation of the RNA and protein molecules. (A) In eukaryotic cells, the mRNA molecule resulting from transcription 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. For convenience, the steps in this figure are depicted as occurring one at a time; in reality, many occur concurrently. For example, the RNA cap is added and splicing begins before transcription has been completed. Because of the coupling between transcription and RNA processing, intact primary transcripts—the full-length RNAs that would, in theory, be produced if no processing had occurred—are found only rarely. (B) In prokaryotes, 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 prokaryotic cells lack a nucleus, transcription and translation take place in a common compartment, and the translation of a bacterial mRNA often begins before its synthesis has been completed.

more difficult to open, the tension should facilitate the partial unwrapping of the DNA in nucleosomes, inasmuch as the release of DNA from the histone core helps to relax this tension.

Any protein that propels itself alone along a DNA strand of a double helix, such as a DNA helicase or an RNA polymerase, tends to generate superhelical tension. In eukaryotes, DNA topoisomerase enzymes rapidly remove this superhelical tension (see pp. 251–253). 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 6–19B). 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 facilitates those genetic processes in bacteria, such as the initiation of transcription by bacterial RNA polymerase, that require helix opening (see Figure 6–11).

Transcription Elongation in Eukaryotes Is Tightly Coupled to RNA Processing

We have seen that bacterial mRNAs are synthesized by the RNA polymerase starting and stopping at specific spots on the genome. The situation in eukaryotes is substantially different. In particular, transcription is only the first of several steps needed to produce a mature mRNA molecule. 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 6–20**).

Both ends of eukaryotic mRNAs are modified: by *capping* on the 5′ end and by *polyadenylation* of the 3′ end (**Figure 6–21**). These special ends allow the cell to assess whether both ends of an mRNA molecule are present (and if the message is therefore intact) before it exports the RNA from the nucleus and translates it

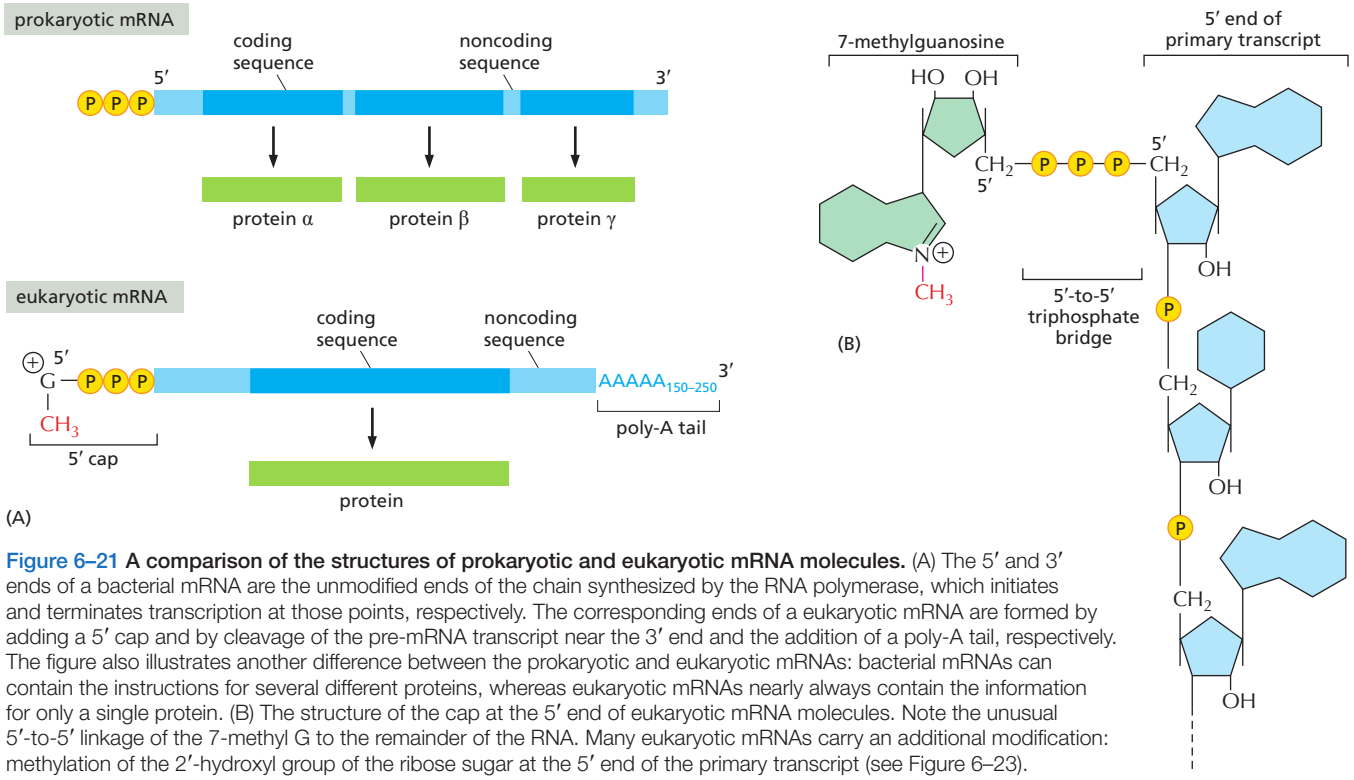


Figure 6–21 A comparison of the structures of prokaryotic and eukaryotic mRNA molecules. (A) The 5′ and 3′ ends of a bacterial mRNA are the unmodified ends of the chain synthesized by the RNA polymerase, which initiates and terminates transcription at those points, respectively. The corresponding ends of a eukaryotic mRNA are formed by adding a 5′ cap and by cleavage of the pre-mRNA transcript near the 3′ end and the addition of a poly-A tail, respectively. The figure also illustrates another difference between the prokaryotic and eukaryotic mRNAs: bacterial mRNAs can contain the instructions for several different proteins, whereas eukaryotic mRNAs nearly always contain the information for only a single protein. (B) The structure of the cap at the 5′ end of eukaryotic mRNA molecules. Note the unusual 5′-to-5′ linkage of the 7-methyl G to the remainder of the RNA. Many eukaryotic mRNAs carry an additional modification: methylation of the 2′-hydroxyl group of the ribose sugar at the 5′ end of the primary transcript (see Figure 6–23).

into protein. RNA splicing joins together the different portions of a protein-coding sequence, and it provides eukaryotes with the ability to synthesize several different proteins from the same gene.

A simple strategy has evolved to couple 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, also called the CTD (C-terminal domain). This phosphorylation, which proceeds gradually as the RNA polymerase initiates transcription and moves along the DNA, 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 are thought 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 not only moves along the DNA synthesizing an RNA molecule, but also processes the RNA that it produces (**Figure 6–22**). Fully extended, the CTD is nearly 10 times longer than the remainder of RNA polymerase. As a flexible protein domain, it serves as a scaffold or tether, holding a variety of proteins close by so that they can rapidly act when needed. This strategy, which greatly speeds up the overall rate of a series of consecutive reactions, is one that is commonly utilized in the cell (see Figures 4–58 and 16–18).

RNA Capping Is the First Modification of Eukaryotic 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 6–21B). 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

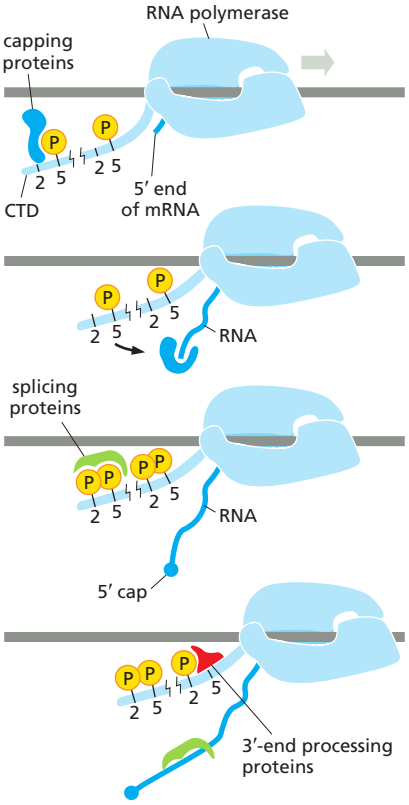


Figure 6–22 Eukaryotic RNA polymerase II as an "RNA factory." As the polymerase transcribes DNA into RNA, it carries RNA-processing proteins on its tail that are transferred to the nascent RNA at the appropriate time. The tail contains 52 tandem repeats of a seven-amino-acid sequence, and there are two 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 6–15). 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 fully dephosphorylated form of RNA polymerase II is competent to begin RNA synthesis at a promoter.

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 6–23**). Because all three enzymes bind to the RNA polymerase tail phosphorylated at the Ser5 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 eukaryotic 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 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 a future mRNA be further 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 eukaryotic genes are typically interrupted by noncoding intervening sequences (introns). Discovered in 1977, this feature of eukaryotic 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, eukaryotic 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 eukaryotic gene is often only a small fraction of the length of the gene (**Figure 6–24**).

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.

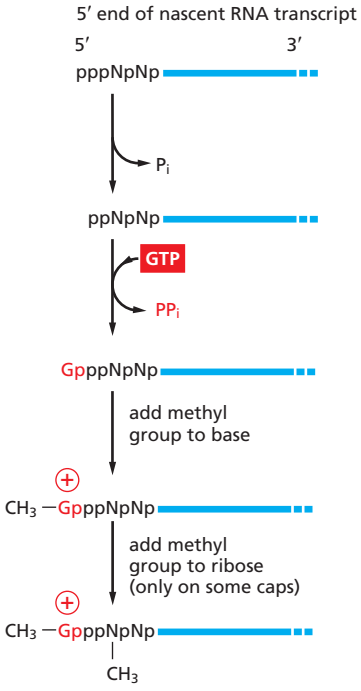
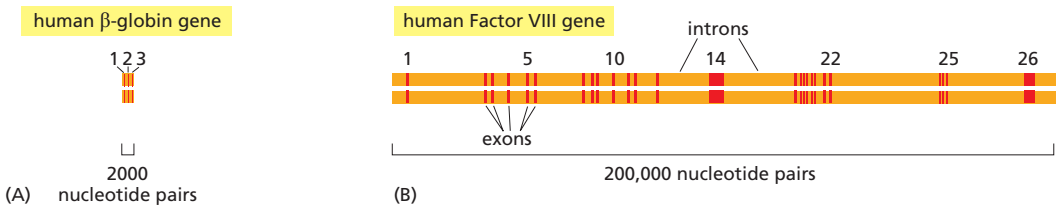


Figure 6–23 The reactions that cap the 5′ end of each RNA molecule synthesized by RNA polymerase II. The final cap contains a novel 5′-to-5′ linkage between the positively charged 7-methyl G residue and the 5′ end of the RNA transcript (see Figure 6–21B). The letter N represents any one of the four ribonucleotides, 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 Wiley-Liss, Inc., a subsidiary of John Wiley & Sons, Inc.)



Each splicing event removes one intron, proceeding through two sequential phosphoryl-transfer reactions known as transesterifications; these join two exons together while removing the intron between them as a "lariat" (Figure 6–25). The machinery that catalyzes pre-mRNA splicing is complex, consisting of five additional RNA molecules and several hundred proteins, and it hydrolyzes many ATP molecules per splicing event. This 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 eukaryotic 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, 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 (see pp. 121–122).

RNA splicing also has a present-day advantage. The transcripts of many eukaryotic genes (estimated at 95% 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 6–26). Rather than being the wasteful process it may have seemed at first sight, RNA splicing enables eukaryotes to increase the 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 6–24 Structure of two human genes showing the arrangement of exons and introns. (A) The relatively small β-globin gene, which encodes a subunit 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.

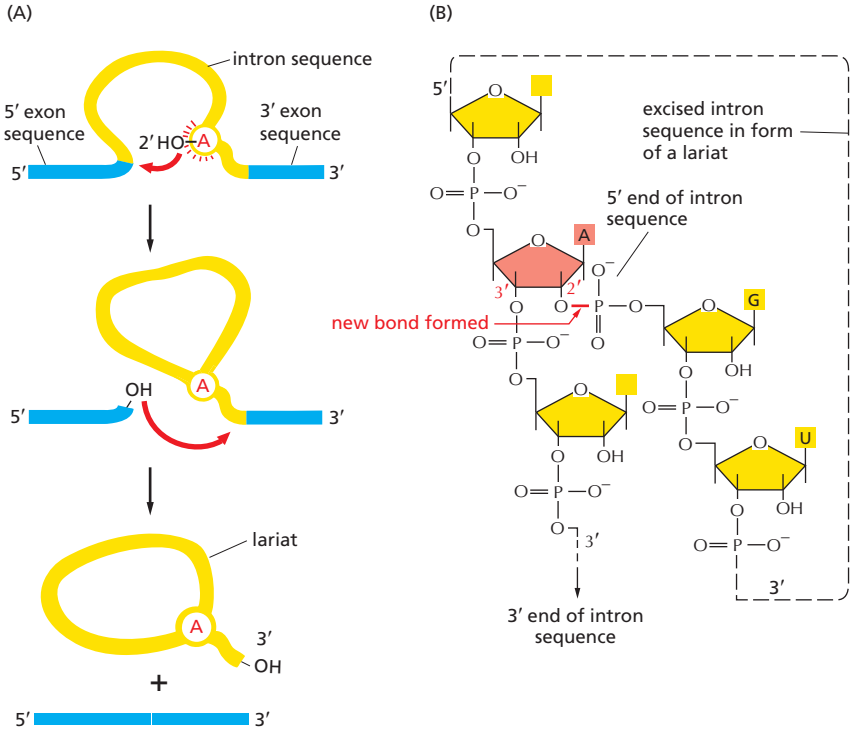


Figure 6–25 The pre-mRNA splicing reaction. (A) In the first step, a specific adenine nucleotide in the intron sequence (indicated in *red*) attacks the 5′ splice site and cuts the sugar-phosphate backbone of the RNA at this point. The cut 5′ end of the intron becomes covalently linked to the adenine 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 broken down into single nucleotides, which are recycled.

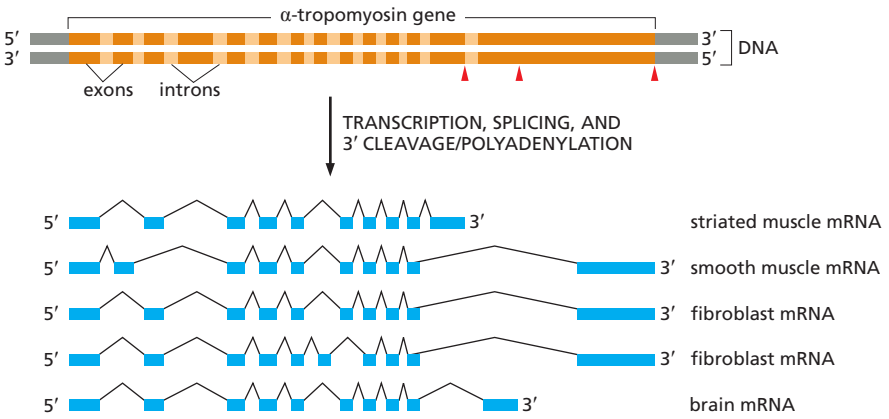


Figure 6–26 Alternative splicing of the α-tropomyosin gene from rat. α-Tropomyosin is a coiled-coil protein (see Figure 3–9) that carries out several tasks, most notably the regulation of contraction in muscle cells. The primary transcript can be spliced in different ways, as indicated in the figure, 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 6–24 requires that the splicing machinery 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 6–27**). However, these consensus sequences are relatively short and can accommodate extensive 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 different proteins produced by 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 catalyze 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. Known as **snRNAs** (**small nuclear RNAs**), each is complexed with at least seven protein subunits to form an snRNP (small nuclear ribonucleoprotein).

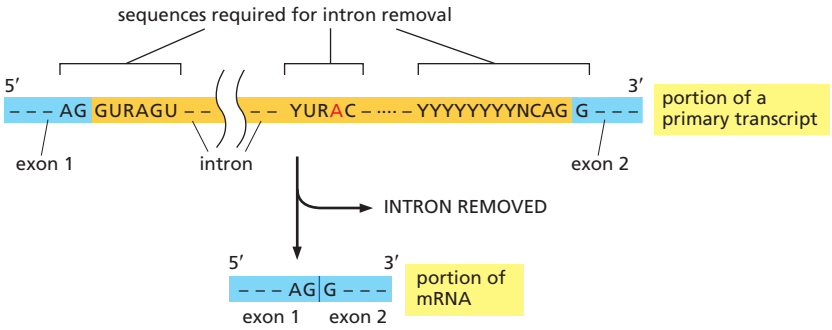
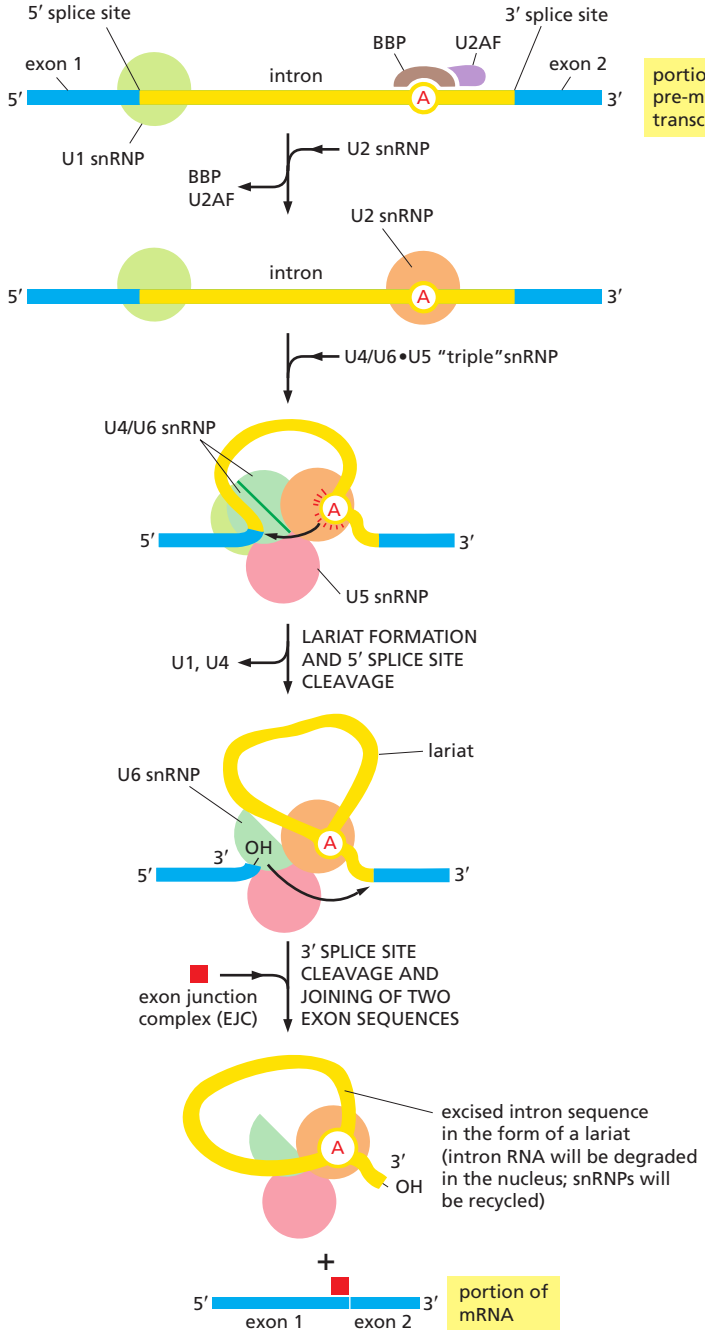


Figure 6–27 The consensus nucleotide sequences in an RNA molecule that signal the beginning and the end of most introns in humans. The three blocks of nucleotide sequences shown are required to remove an intron sequence. Here A, G, U, and C are the standard RNA nucleotides; R stands for purines (A or G); and Y stands for pyrimidines (C or U). The A highlighted in *red* forms the branch point of the lariat produced by splicing (see Figure 6–25). 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, 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.

These snRNPs form the core of the **spliceosome**, the large assembly of RNA and protein molecules that performs pre-mRNA splicing in the cell. 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.

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 and those that have already performed their tasks are jettisoned (**Figure 6–28**). 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.



The U1 snRNP forms base pairs with the 5′ splice junction (see Figure 6–29) 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.

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 break apart the U4/U6 base pairs, allowing U6 to displace U1 at the 5′ splice junction (see Figure 6–29). This creates the active site that catalyzes the first phosphoryl-transferase reaction.

Additional RNA–RNA rearrangements create the active site for the second phosphoryl-transferase reaction, which then completes the splice (see Figure 6–25A).

Figure 6–28 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 required (see Figure 6–25A and **Movie 6.5**). As indicated, a set of proteins called the exon junction complex (EJC) remains on the spliced mRNA molecule; its subsequent role will be discussed shortly.

The Spliceosome Uses ATP Hydrolysis to Produce a Complex Series of RNA–RNA Rearrangements

ATP hydrolysis is not required for the chemistry of RNA splicing *per se* since the two transesterification reactions preserve the high-energy phosphate bonds. However, extensive ATP hydrolysis 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. Each successful splice requires approximately 200 proteins, if we include those that form the snRNPs.

What is the purpose of these rearrangements? First, they allow the splicing signals on the pre-RNA to be examined by snRNPs several times during the course of splicing. For example, the U1 snRNP initially recognizes the 5′ splice site through conventional base-pairing; as splicing proceeds, these base pairs are broken (using the energy of ATP hydrolysis) and U1 is replaced by U6 (**Figure 6–29**). This type of RNA–RNA rearrangement (in which the formation of one RNA–RNA interaction requires the disruption of another) occurs several times during splicing and allows the spliceosomes to check and recheck the splicing signals, thereby increasing the overall accuracy of splicing. Second, the rearrangements that take place in the spliceosome create the active sites for the two transesterification reactions. These two active sites are created, one after the other, and only after the splicing signals on the pre-mRNA have been checked several times. This orderly progression ensures that splicing accidents occur only rarely.

One of the most surprising features of the spliceosome is the nature of the catalytic sites: they are formed by both protein and RNA molecules, although the RNA molecules catalyze the actual chemistry of splicing. In the last section of this chapter, we discuss in general terms the structural and chemical properties of RNA molecules that allow them to act as catalysts.

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 frequent splicing mistakes—such as exon skipping and the use of "cryptic" splice sites (**Figure 6–30**). The fidelity mechanisms built into the spliceosome to suppress errors, however, are supplemented by two additional strategies that further increase the accuracy of splicing. The first is a simple consequence of splicing being coupled to transcription. As transcription proceeds, the phosphorylated tail of RNA polymerase carries several components of the spliceosome (see Figure

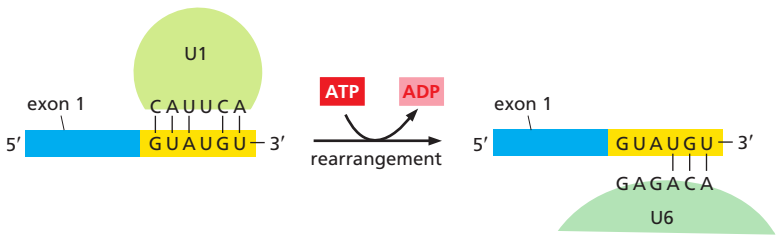


Figure 6–29 One of the many rearrangements that take place in the spliceosome during pre-mRNA splicing. This example comes from the yeast *Saccharomyces cerevisiae*, in which the nucleotide sequences involved are slightly different from those in human cells. The exchange of U1 snRNP for U6 snRNP occurs just before the first phosphoryl-transfer reaction (see Figure 6–28). 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.

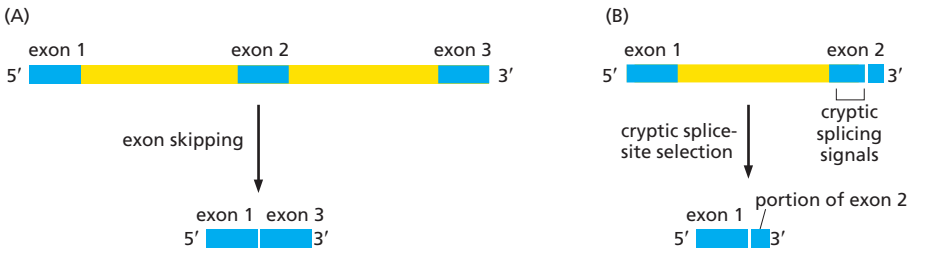


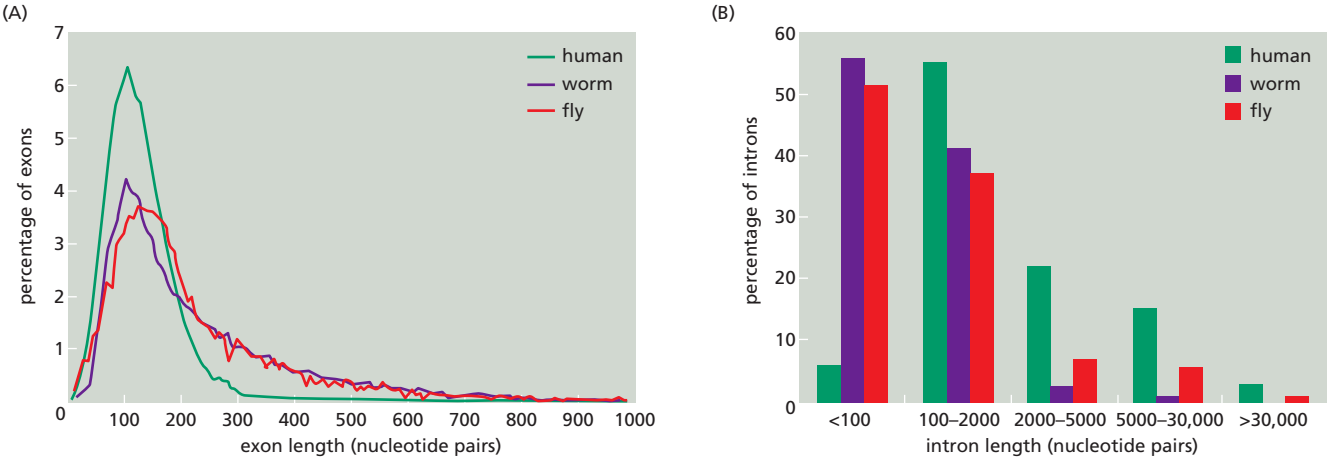
Figure 6–30 Two types of splicing errors. (A) Exon skipping. (B) Cryptic splice-site selection. Cryptic splicing signals are nucleotide sequences of RNA that closely resemble true splicing signals and are sometimes mistakenly used by the spliceosome.

6–22), and these components are transferred directly from the polymerase to the RNA as the RNA emerges from the polymerase. 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 only with the single 3′ splice site that emerges next from the polymerase; the potential 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" also helps 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 eukaryotic organisms (**Figure 6–31**). Through exon definition, the splicing machinery can seek out the relatively homogeneously 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 6–32**). These proteins, in turn, recruit U1 snRNA, which marks the downstream exon boundary, and U2 snRNA, 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 avoids "near miss" 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 evolve 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 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.

Figure 6–31 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.)



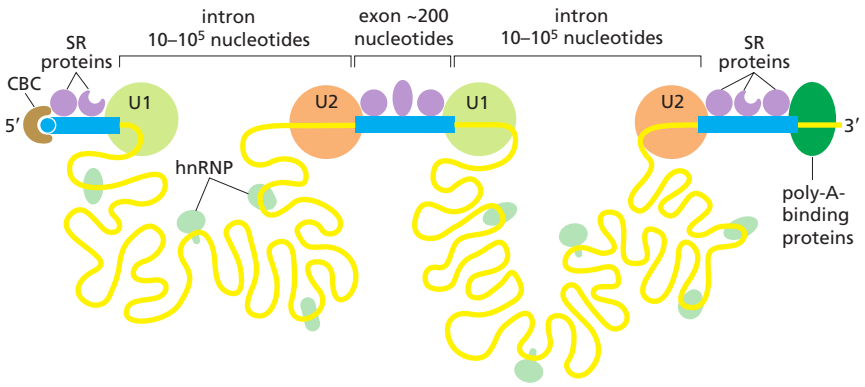


Figure 6–32 The exon definition hypothesis. According to this idea, SR proteins bind to each exon sequence in the pre-mRNA and thereby help to guide the snRNPs to the proper intron/exon boundaries. This demarcation of exons by the SR proteins occurs co-transcriptionally, beginning at the CBC (cap-binding complex) at the 5′ end. It has been proposed that a group of proteins known as the heterogeneous nuclear ribonucleoproteins (hnRNPs) may preferentially associate with intron sequences, further helping the spliceosome distinguish introns from exons. (Adapted from R. Reed, *Curr. Opin. Cell Biol.* 12:340–345, 2000. With permission from Elsevier.)

Chromatin Structure Affects RNA Splicing

Although it may seem at first counterintuitive, the way a gene is packaged into chromatin can affect how the RNA transcript of that gene is ultimately spliced. Nucleosomes tend to be positioned over exons (which are, on average, close to the length of DNA in a nucleosome), and it has been proposed that these act as "speed bumps," allowing the proteins responsible for exon definition to assemble on the RNA as it emerges from the polymerase. In addition, changes in chromatin structure are used to alter splicing patterns. There are two ways this can happen. First, because splicing and transcription are coupled, the rate at which RNA polymerase moves along DNA can affect RNA splicing. For example, if polymerase is moving slowly, exon skipping (see Figure 6–30A) is minimized: assembly of the initial spliceosome may be complete before an alternative choice of splice site even emerges from the RNA polymerase. The nucleosomes in condensed chromatin can cause polymerase to pause; the pattern of pauses in turn affects the extent of RNA exposed at any given time to the splicing machinery.

There is a second and more direct way that chromatin structure can affect RNA splicing. Although the details are not yet understood, specific histone modifications attract components of the spliceosome, and, because the chromatin being transcribed is in close association with the nascent RNA, these splicing components can easily be transferred to the emerging RNA. In this way, certain types of histone modifications can affect the final pattern of splicing.

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 strength 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, chromatin structure, and the "bookkeeping" that underlies exon definition. We do not know exactly 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.

Thus, 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 6–33**). Most commonly, an exon is simply skipped (Figure 6–33B). In other cases, the mutation causes a cryptic splice junction to be efficiently used (Figure 6–33C). 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 inherent plasticity in the process of RNA splicing suggests that changes in splicing patterns caused by random mutations have been important in the evolution of genes and organisms. It also means that mutations that affect splicing can be severely detrimental to the organism: in addition to the β thalassemia, example presented in Figure 6–33, aberrant

splicing plays important roles in the development of cystic fibrosis, frontotemporal dementia, Parkinson's disease, retinitis pigmentosa, spinal muscular atrophy, myotonic dystrophy, premature aging, and cancer. It has been estimated that of the many point mutations that cause inherited human diseases, 10% produce aberrant splicing of the gene containing the mutation.

The plasticity of RNA splicing also means that the cell can easily 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 and that this is a common strategy to enhance the coding potential of genomes. 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 6–26). 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 has 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 critical 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. In these cases, the RNA molecule folds into a specific three-dimensional structure that brings the intron/exon junctions together and catalyzes the two transesterification reactions. 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. Because the basic chemistry of some self-splicing reactions is so similar to pre-mRNA splicing, it has been proposed that the much more involved process of pre-mRNA splicing evolved from a simpler, ancestral form of RNA self-splicing.

RNA-Processing Enzymes Generate the 3′ End of Eukaryotic mRNAs

We have seen that 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. In this section, we shall see 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.

The position of the 3′ end of each mRNA molecule is specified by signals encoded in the genome (**Figure 6–34**). 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 6–35**). Two multisubunit proteins, called CstF (cleavage stimulation factor) and CPSF (cleavage and polyadenylation specificity factor), are of special importance.

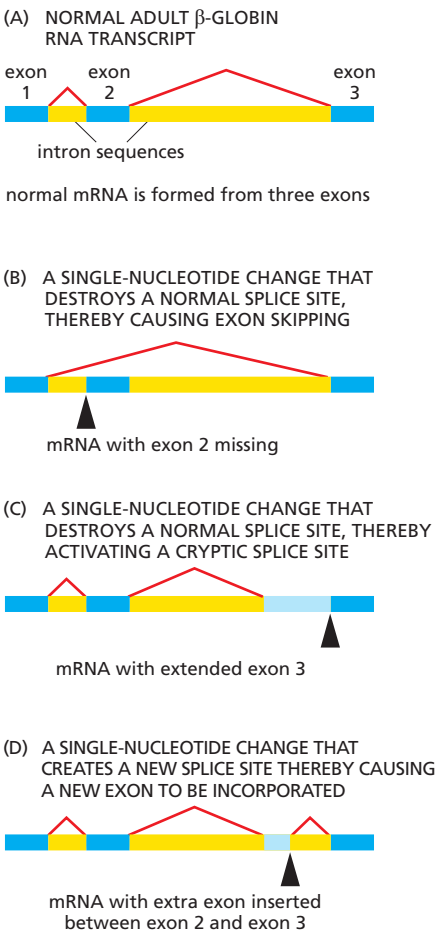


Figure 6–33 Abnormal processing of the β-globin primary RNA transcript in humans with the disease β thalassemia. In the examples shown, the disease (a severe anemia due to aberrant hemoglobin synthesis) is caused by splice-site mutations found in the genomes of affected patients. The *dark blue boxes* represent the three normal exon sequences; the *red lines* connect the 5′ and 3′ splice sites that are used. In (B), (C), and (D), the *light blue boxes* depict new nucleotide sequences included in the final mRNA molecule as a result of the mutation denoted by the *black arrowhead*. Note that when a mutation leaves a normal splice site without a partner, an exon is skipped (B) or one or more abnormal cryptic splice sites nearby is used as the partner site (C). [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.]

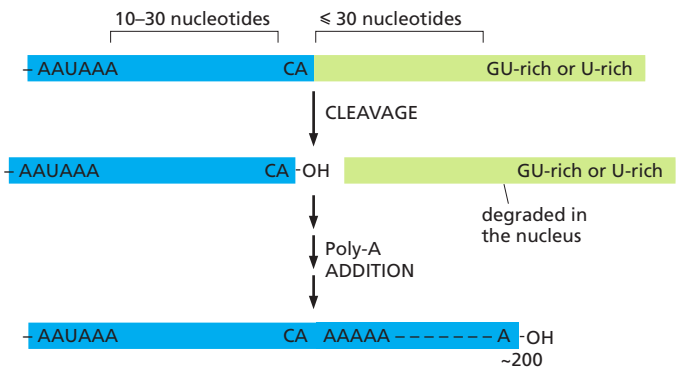


Figure 6–34 Consensus nucleotide sequences that direct cleavage and polyadenylation to form the 3′ end of a eukaryotic mRNA. These sequences are encoded in the genome, and specific proteins recognize them—as RNA—after they are transcribed. As shown in Figure 6–35, the hexamer AAUAAA is bound by CPSF and the GU-rich element beyond the cleavage site is bound by CstF; the CA sequence is bound by a third protein 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.

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 their recognition sequences on the emerging RNA molecule, additional proteins assemble with them to create the 3′ end of the mRNA. First, the RNA is cleaved from the polymerase (see Figure 6–35). 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. But unlike other RNA polymerases, poly-A polymerase does not require a template; hence the poly-A tail of eukaryotic mRNAs 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, help determine the final length of the tail.

After the 3′-end of a eukaryotic pre-mRNA molecule has been cleaved, the RNA polymerase II continues to transcribe, in some cases for hundreds of nucleotides. Once 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 carried along on the polymerase tail. Apparently, it is this continued RNA degradation that eventually causes the RNA polymerase to release its grip on the template and terminate transcription.

Mature Eukaryotic mRNAs Are Selectively Exported from the Nucleus

Eukaryotic pre-mRNA synthesis and processing take place in an orderly fashion within the cell nucleus. But of the pre-mRNA that is synthesized, only a small fraction—the mature mRNA—is of further use to the cell. Most of the rest—excised introns, broken RNAs, and aberrantly processed pre-mRNAs—is not only useless but potentially dangerous. How does the cell distinguish between the relatively rare mature mRNA molecules it wishes to keep and the overwhelming amount of debris created by RNA processing?

The answer is that, as an RNA molecule is processed, it loses certain proteins and acquires others. For example, we have seen that acquisition of cap-binding complexes, exon junction complexes, and 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 an snRNP protein 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

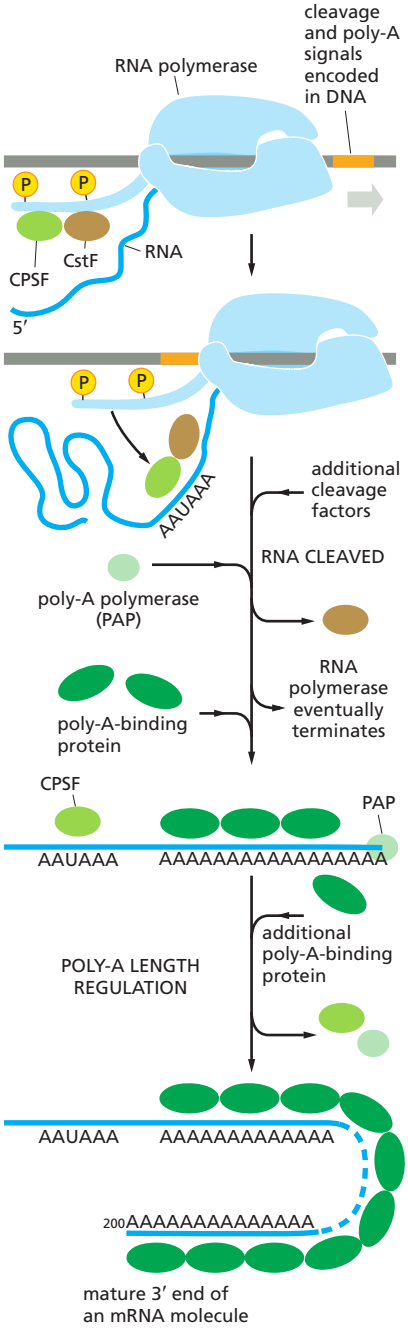


Figure 6–35 Some of the major steps in generating the 3′ end of a eukaryotic mRNA. This process is much more complicated than the analogous process in bacteria, where the RNA polymerase simply stops at a termination signal and releases both the 3′ end of its transcript and the DNA template (see Figure 6–11).

and other RNA debris (excised intron sequences, for example) 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 (Figure 6–36). Eukaryotic 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). Some of these proteins (there are approximately 30 different ones in humans) unwind the hairpin helices in 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 typical in complex organisms (see Figure 6–31) and these may 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 6–37). Small molecules (less than 60,000 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 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, in many 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 is then used to export a new mRNA molecule.

The export of mRNA–protein complexes from the nucleus can be readily 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 6–37). 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 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 (Figure 6–38).

The analysis just described has been complemented by new methods that allow researchers to track the fate of more typical mRNA molecules, which can

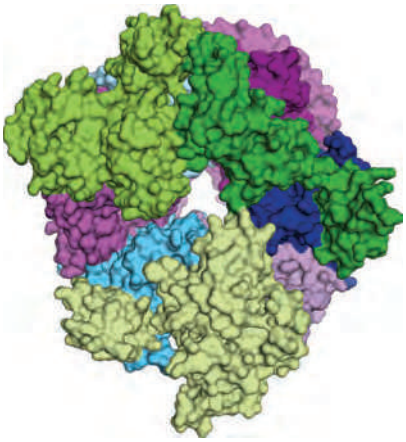
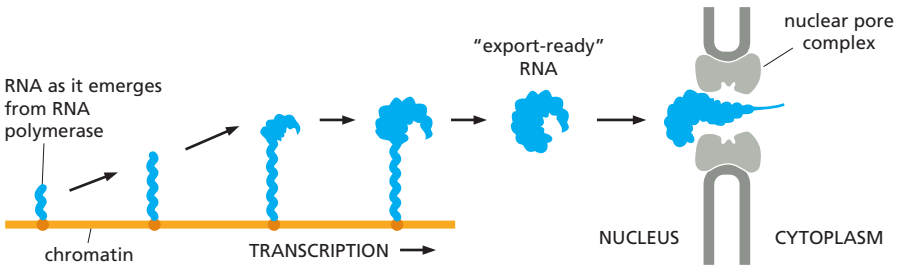
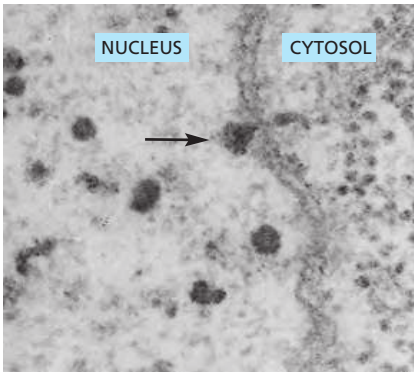


Figure 6–36 Structure of the core of human RNA exosome. RNA is fed into one end of the central pore and is degraded by RNAses that associate with the other end. Nine different protein subunits (each represented by a different color) make up this large ring structure. Eukaryotic cells have both a nuclear exosome and a cytoplasmic exosome; both forms include the core exosome shown here and additional subunits (including specialized RNAses) that differentiate the two forms. The nuclear exosome degrades aberrant RNAs before they are exported to the cytosol. It also processes certain types of RNA (for example, the ribosomal RNAs) to produce their final form. The cytoplasmic form of the exosome is responsible for degrading mRNAs in the cytosol, and is thus crucial in determining the lifetime of each mRNA molecule. (PDB code: 2NN6.)



(A)



(B)

Figure 6–37 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 large and abundant insect RNA, called the Balbiani Ring mRNA, is based on electron microscope micrographs such as that shown in (B). (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.)

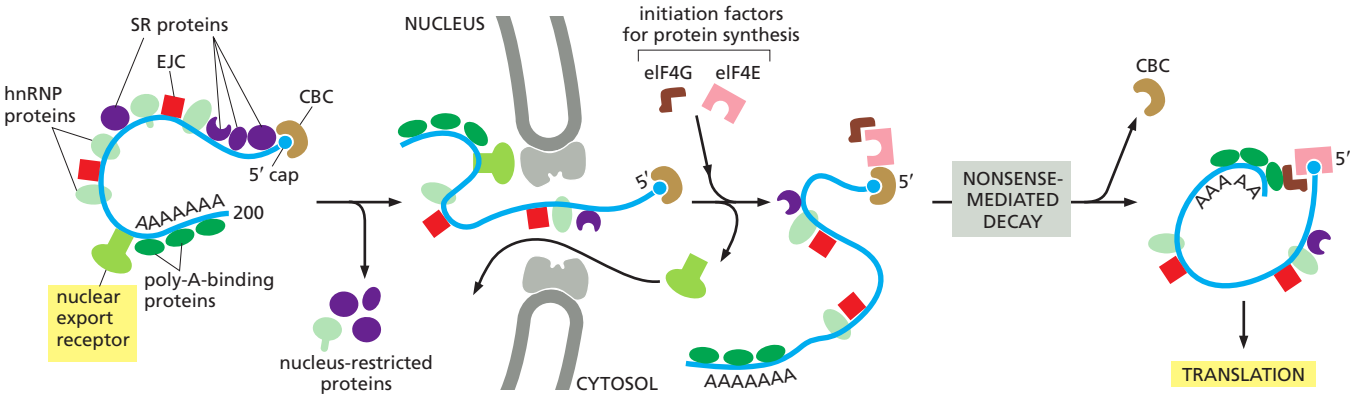


Figure 6–38 **Schematic illustration of an export-ready mRNA molecule and its transport through the nuclear pore.** As indicated, some proteins travel with the mRNA as it moves through the pore, whereas others remain in the nucleus. The nuclear export receptor for mRNAs is a complex of proteins that binds to an mRNA molecule once it has been correctly spliced and polyadenylated. After the mRNA has been exported to the cytosol, this export receptor dissociates from the mRNA and is re-imported into the nucleus, where it can be used again. The final check indicated here, called *nonsense-mediated decay*, will be described later in the chapter.

be fluorescently labeled and observed individually. A typical RNA molecule is released from its site of transcription and spends several minutes diffusing to a nuclear pore complex. During this time it is likely that RNA processing events continue and that the RNA sheds previously bound proteins and acquires new ones. Once it arrives at the entrance to the pore, the "export-ready" mRNA hovers for several seconds, during which time the completion of processing may occur, and then is transported through the pore very rapidly, in tens of milliseconds. Some mRNA–protein complexes are very large, and how they move through the nuclear pore complexes so rapidly remains a mystery.

Some of the 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.

But before discussing what happens to mRNAs in the cytosol, we briefly consider how the synthesis and processing of some noncoding RNA molecules occurs. There are many types of noncoding RNAs produced by cells (see Table 6–1, p. 305), but here we focus on the rRNAs, which are critically important for the translation of mRNAs into protein.

Noncoding RNAs Are Also Synthesized and Processed in the Nucleus

Only a few percent of the dry weight of a mammalian cell is RNA; of that, only about 3–5% is mRNA. The bulk of the RNA in cells performs structural and catalytic functions (see Table 6–1). 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—eukaryotes 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.

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 (see Figure 6–3). 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

Figure 6–39 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 6–10. (From V.E. Foe, *Cold Spring Harb. Symp. Quant. Biol.* 42:723–740, 1978. With permission from Cold Spring Harbor Laboratory Press.)

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 6–39**).

There are four types of eukaryotic 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 6–40**); the fourth (5S 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 6–41**A). The functions of these modifications are not understood in detail, but they probably aid in the folding and assembly of the final rRNAs, or subtly alter the function of ribosomes. Each modification is made at a specific position in the precursor rRNA, specified by "guide RNAs," which position themselves on the precursor rRNA through base-pairing and thereby bring an RNA-modifying enzyme to the appropriate position (Figure 6–41B). 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

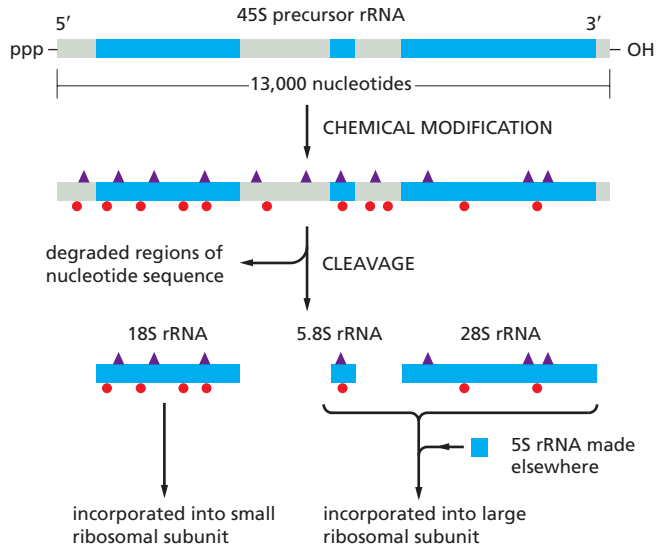


Figure 6–40 The chemical modification and nucleolytic processing of a eukaryotic 45S precursor rRNA molecule into three separate ribosomal RNAs. Two types of chemical modifications (color-coded as indicated in Figure 6–41) are made to the precursor rRNA before it is cleaved. Nearly half of the nucleotide sequences in this precursor rRNA are discarded and degraded in the nucleus by the exosome. 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.

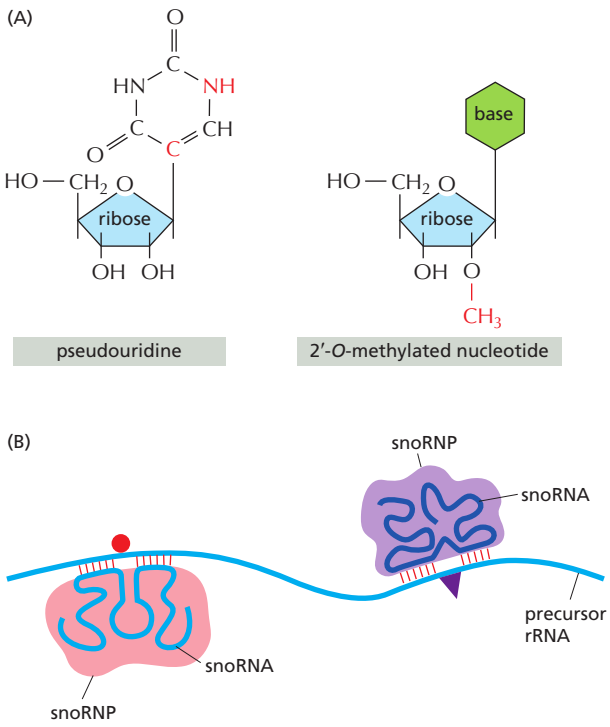


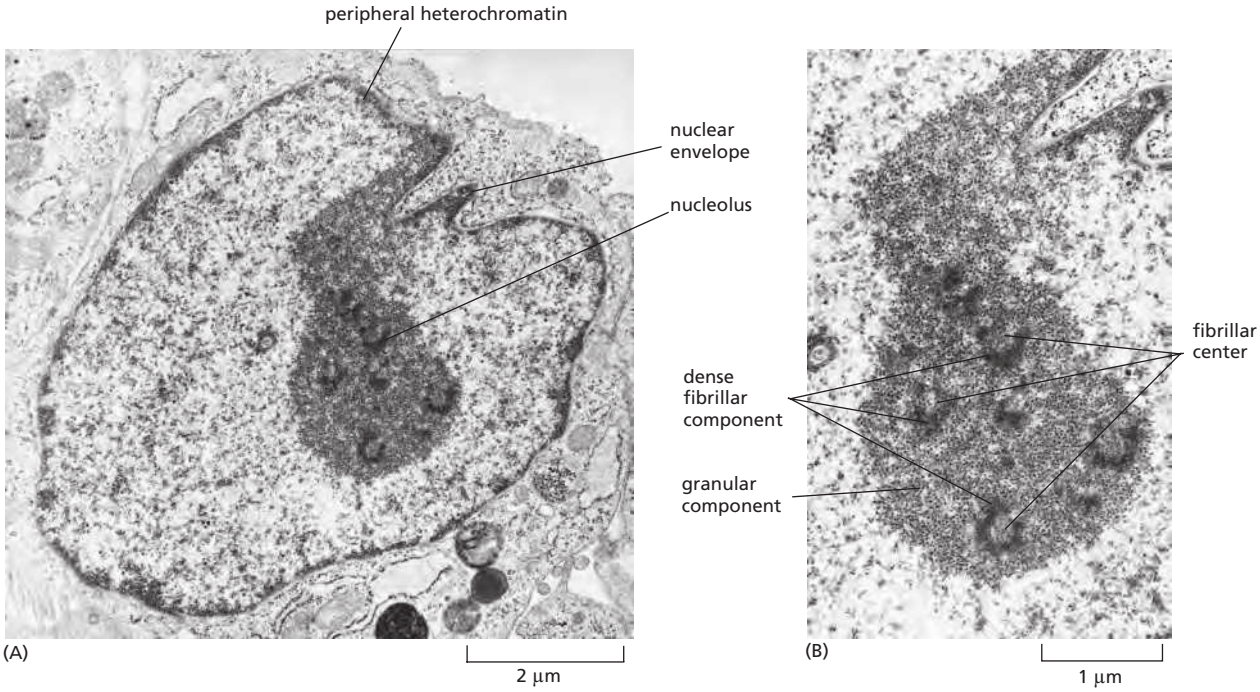
Figure 6–41 Modifications of the precursor rRNA by guide RNAs. (A) Two prominent covalent modifications made to rRNA; the differences from the initially incorporated nucleotide are indicated by *red* atoms. Pseudouridine is an isomer of uridine; the base has been "rotated," and is attached to the *red* C rather than to the *red* N of the sugar (compare to Figure 6–5B). (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 (small nucleolar ribonucleoproteins). snoRNPs contain both the guide sequences and the enzymes that modify the rRNA.

the introns of other genes, especially those encoding ribosomal proteins. They are synthesized by RNA polymerase II and processed from excised intron sequences.

The Nucleolus Is a Ribosome-Producing Factory

The nucleolus is the most obvious structure seen in the nucleus of a eukaryotic cell when viewed in the light microscope. 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 6–42**); instead, it is a huge aggregate

Figure 6–42 Electron micrograph of a thin section of a nucleolus in a human fibroblast, showing its three distinct zones. (A) View of entire nucleus. (B) Higher-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.)



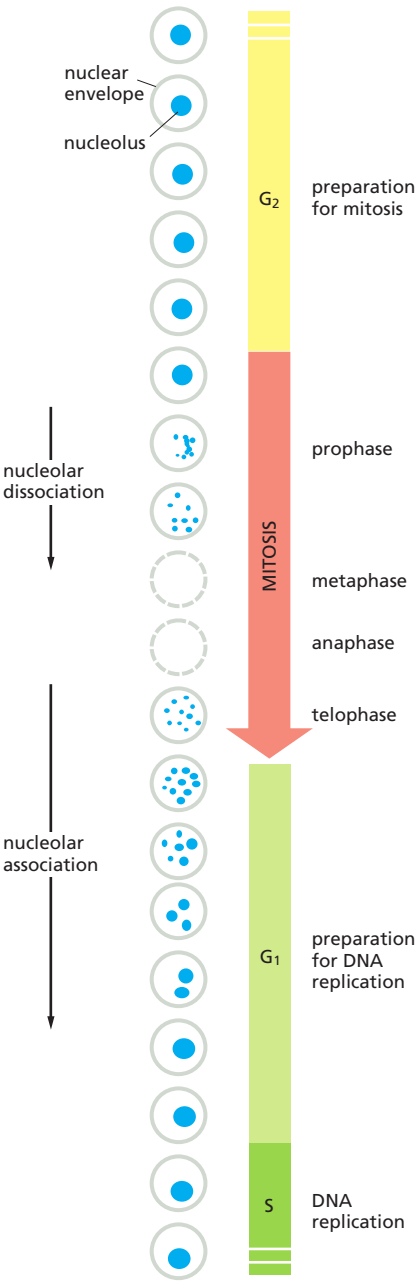


Figure 6–43 Changes in the appearance of the nucleolus in a human cell during the cell cycle. Only the cell nucleus is represented in this diagram. In most eukaryotic cells, the nuclear envelope breaks down during mitosis, as indicated by the *dashed circles*.

of macromolecules, including the rRNA genes themselves, precursor rRNAs, mature rRNAs, rRNA-processing enzymes, snoRNPs, a large set of assembly factors (including ATPases, GTPases, protein kinases, and RNA helicases), ribosomal proteins, and partly assembled ribosomes. The close association of all these components allows the assembly of ribosomes to occur rapidly and smoothly.

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 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 fragments and then disappears. Then, in the telophase part of mitosis, as chromosomes return to their semi-dispersed state, the tips of the 10 chromosomes reform small nucleoli, which progressively coalesce into a single nucleolus (**Figure 6–43** and **Figure 6–44**). 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 6–45**. In addition to its central role in ribosome biogenesis, the nucleolus is the site where other noncoding RNAs are produced and other RNA–protein complexes are assembled. For example, the U6 snRNP, which functions in pre-mRNA splicing (see Figure 6–28), 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 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, the genes encoding tRNAs are clustered in the nucleolus. Thus, the nucleolus can be thought of as a large factory at which different noncoding RNAs are transcribed, processed, and assembled with proteins to form a large variety of ribonucleoprotein complexes.

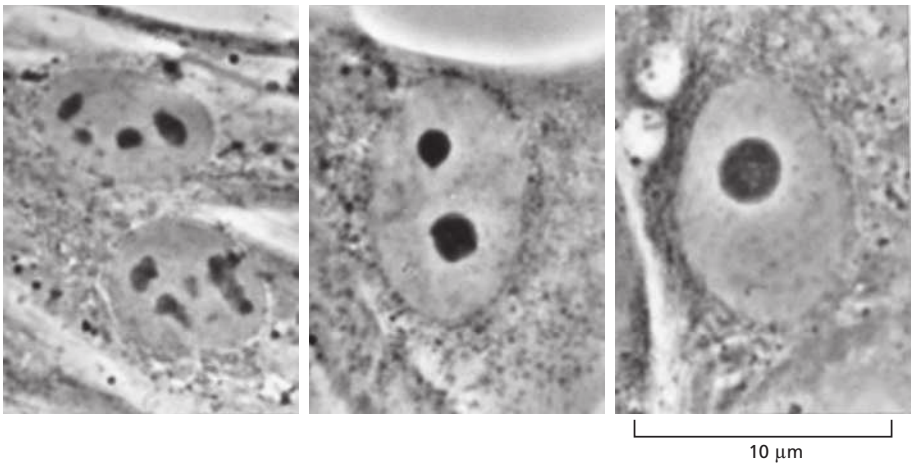


Figure 6–44 Nucleolar fusion. These light micrographs of human fibroblasts grown in culture show various stages of nucleolar fusion. After mitosis, each of the 10 human chromosomes that carry a cluster of rRNA genes begins to form a tiny nucleolus, but these rapidly coalesce as they grow to form the single large nucleolus typical of many interphase cells. (Courtesy of E.G. Jordan and J. McGovern.)

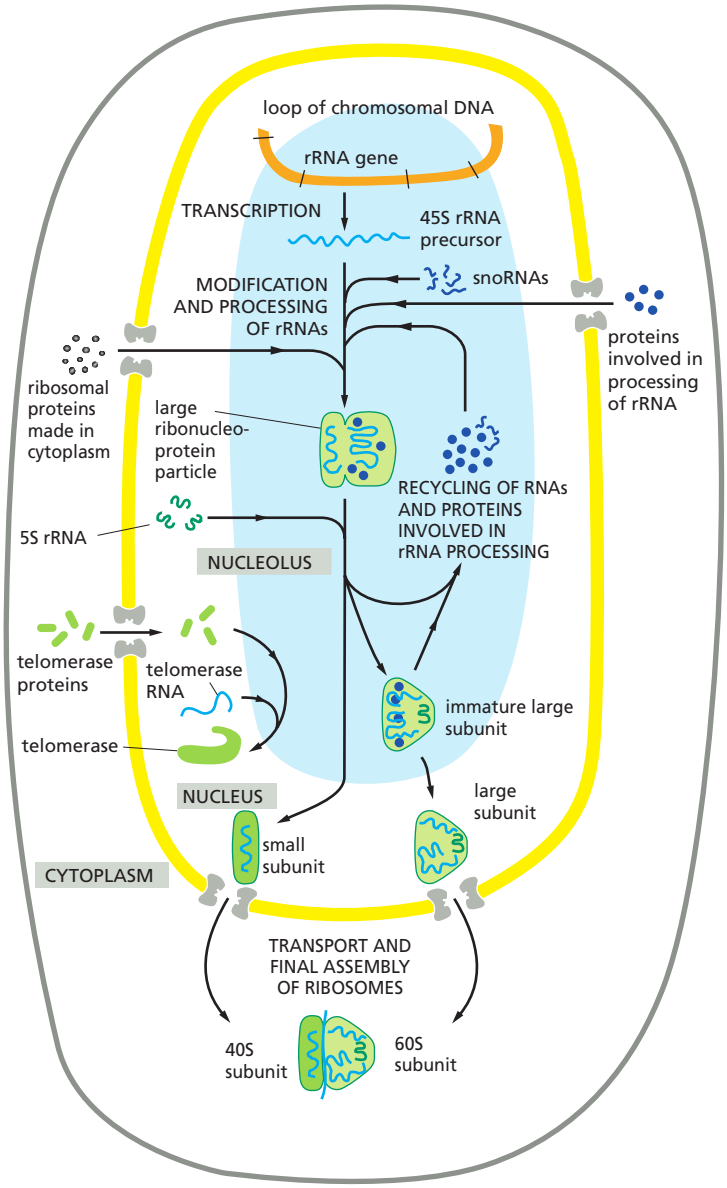


Figure 6–45 The function of the nucleolus in ribosome and other ribonucleoprotein synthesis. The 45S precursor rRNA is packaged in a large ribonucleoprotein particle containing many ribosomal proteins imported from the cytoplasm. While this particle remains at the nucleolus, selected components are added and others discarded as it is processed into immature large and small ribosomal subunits. The two ribosomal subunits attain their final functional form only after each is individually transported through the nuclear pores into the cytoplasm. Other ribonucleoprotein complexes, including telomerase shown here, are also assembled in the nucleolus.

The Nucleus Contains a Variety of Subnuclear Aggregates

Although the nucleolus is the most prominent structure in the nucleus, several other nuclear bodies have been observed and studied (**Figure 6–46**). These include Cajal bodies (named for the scientist who first described them in 1906) and interchromatin granule clusters (also called "speckles"). Like the nucleolus, these other nuclear structures lack membranes and are highly dynamic depending on the needs of the cell. Their assembly is likely mediated by the association of low complexity protein domains, as described in Chapter 3 (see Figure 3–36). Their appearance is 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 are sites where the snRNPs and snoRNPs undergo their final maturation steps, and where the snRNPs are recycled and their RNAs are "reset" after the rearrangements that occur during splicing (see p. 321). 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.

Scientists have had difficulties in working out the function of these small subnuclear structures, in part because their appearances can change dramatically as cells traverse the cell cycle or respond to changes in their environment. Moreover,

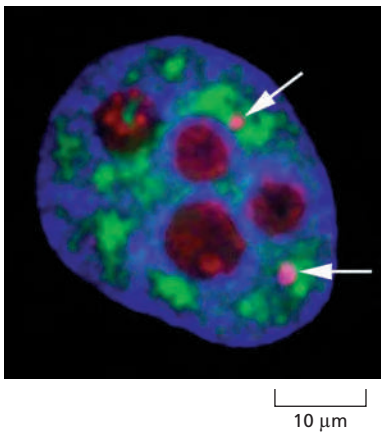


Figure 6–46 Visualization of some prominent nuclear bodies. The protein fibrillarin *(red)*, a component of several snoRNPs, is present at both nucleoli and Cajal bodies; the latter are indicated by the arrows. The Cajal bodies (but not the nucleoli) are also highlighted by staining one of their main components, the protein coilin; the superposition of the snoRNP and coilin stains appears *pink*. Interchromatin granule clusters *(green)* have been revealed by using antibodies against a protein involved in pre-mRNA splicing. DNA is stained blue by the dye DAPI. (From J.R. Swedlow and A.I. Lamond, *Gen. Biol.* 2:1–7, 2001. With permission from BioMed Central. Micrograph courtesy of Judith Sleeman.)

disrupting a particular type of nuclear body often has little effect on cell viability. It seems that the main function of these aggregates is to bring components together at high concentration in order to speed up their assembly. For example, it is estimated that assembly of the U4/U6 snRNP (see Figure 6–28) occurs ten times more rapidly in Cajal bodies than would be the case if the same number of components were dispersed throughout the nucleus. Consequently, Cajal bodies appear dispensible in many types of cells but are absolutely required in situations where cells must proliferate rapidly, such as in early vertebrate development. Here, protein synthesis (which depends on RNA splicing) must be especially rapid, and delays can be lethal.

Given the prominence of nuclear bodies in RNA processing, it might be expected that pre-mRNA splicing would occur in a particular location in the nucleus, as it requires numerous RNA and protein components. However, as we have seen, 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 takes place in only several thousand sites in the nucleus. These sites are highly dynamic and probably result from the association of transcription and splicing components to create small *factories*, the name given to specific aggregates containing a high local concentration of selected components that create biochemical assembly lines (**Figure 6–47**). Interchromatin

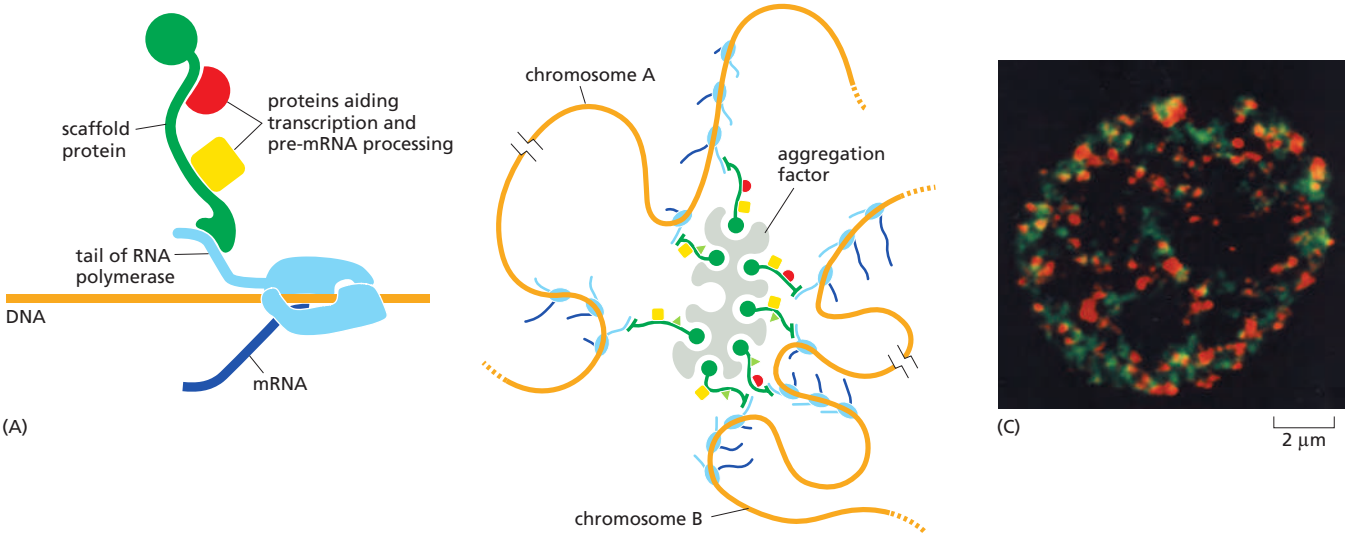


Figure 6–47 A model for an mRNA production factory. mRNA production is made more efficient in the nucleus by an aggregation of the many components needed for transcription and pre-mRNA processing, thereby producing a specialized biochemical factory. In (A), a postulated scaffold protein holds various components in the proximity of a transcribing RNA polymerase. Other key components are bound directly to the RNA polymerase tail, which likewise serves as a scaffold (see Figure 6–22), but for simplicity these are not shown here. In (B), a large number of such scaffolds have been brought together to form an aggregate that is highly enriched in the many components needed for the synthesis and processing of pre-mRNAs. Such a scaffold model can account for the several thousand sites of active RNA transcription and processing typically observed in the nucleus of a mammalian cell, each of which has a diameter of roughly 100nm and is estimated to contain, on average, about 10 RNA polymerase II molecules in addition to many other proteins. (C) Here, mRNA production factories and DNA replication factories have been visualized in the same mammalian cell by briefly incorporating differently modified nucleotides into each nucleic acid and detecting the RNA and DNA produced using antibodies, one *(green)* detecting the newly synthesized DNA and the other *(red)* detecting the newly synthesized RNA. (C, from D.G. Wansink et al., *J. Cell Sci.* 107:1449–1456, 1994. With permission from The Company of Biologists.)

granule clusters—which contain stockpiles of RNA processing components—are often observed next to these sites of transcription, as though poised to replenish supplies. We can thus view the nucleus as organized into subdomains, with snRNPs, snoRNPs, and other nuclear components moving among them in an orderly fashion according to the needs of the cell.

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 after 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 eukaryotic 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 eukaryotic mRNA. This enzyme requires a set of additional proteins, both the general transcription factors, and specific transcriptional activator proteins, to initiate transcription on a DNA template. It requires 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. Only properly processed mRNAs are passed through nuclear pore complexes into the cytosol, where they are translated into protein.

For many genes, RNA, rather than protein, is the final product. In eukaryotes, 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. The high concentration of components in such "factories" ensures that the processes being catalyzed are rapid and efficient.

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 the RNAs 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 strong focus of attention for 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. 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 was the code 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.

GCA GCC GCG GCU	AGA AGG CGA CGC CGG CGU	GAC GAU	AAC AAU	UGC UGU	GAA GAG	CAA CAG	GGA GGC GGG GGU	CAC CAU	AUA AUC AUU	UUA UUG CUA CUC CUG CUU	AAA AAG	AUG	UUC UUU	CCA CCC CCG CCU	AGC AGU UCA UCC UCG UCU	ACA ACC ACG ACU	UGG	UAC UAU	GUA GUC GUG GUU	UAA UAG 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	

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 instead 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 6–48**. 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 tiny genomes have been able to accommodate minor changes to the code (discussed in Chapter 14).

In principle, an RNA sequence can be translated in any one of three different **reading frames**, depending on where the decoding process begins (**Figure 6–49**). 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.

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.

Figure 6–48 The genetic code. The standard one-letter abbreviation for each amino acid is presented below its three-letter abbreviation (see Panel 3–1, pp. 112–113, 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.

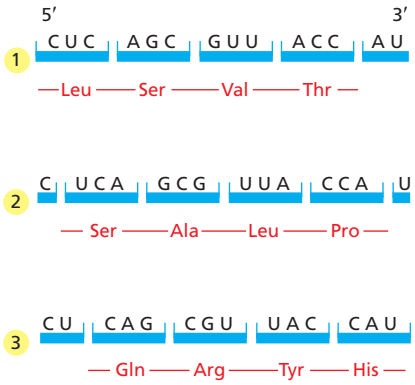


Figure 6–49 The three possible reading frames in protein synthesis. In the process of translating a nucleotide sequence *(blue)* into an amino acid sequence *(red),* the sequence of nucleotides in an mRNA molecule is read from the 5′ end to the 3′ end in consecutive sets of three nucleotides. In principle, 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.

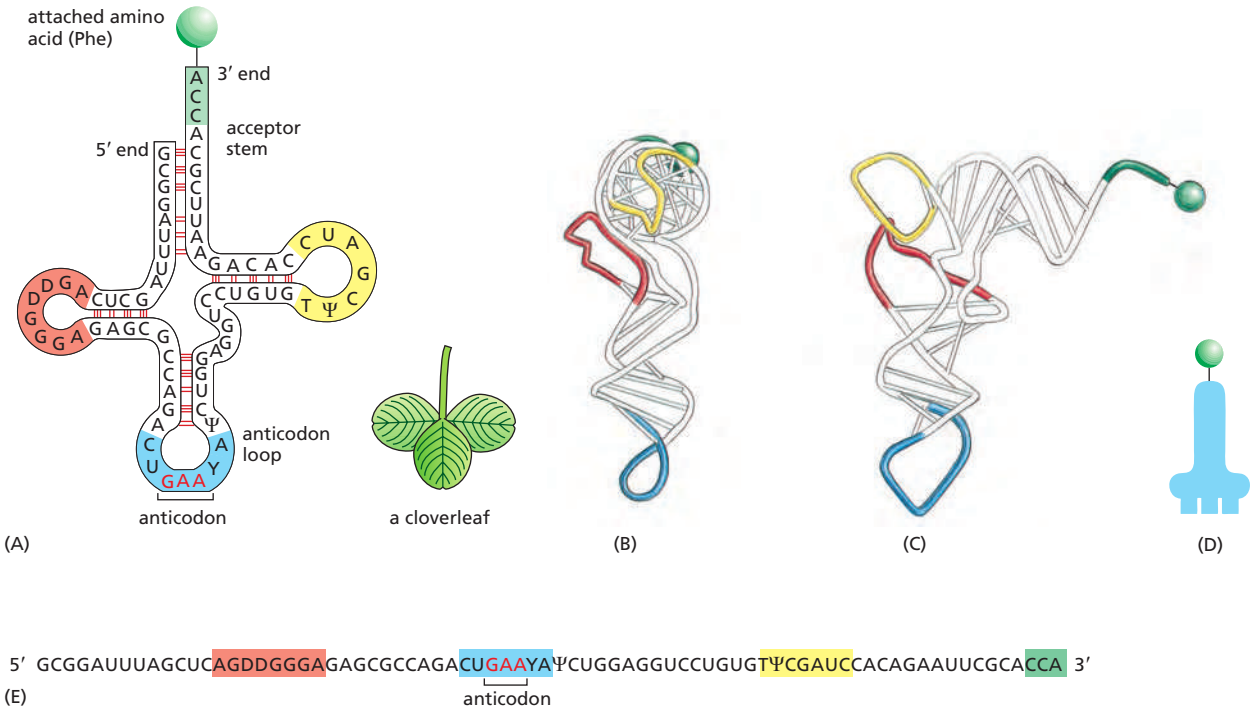
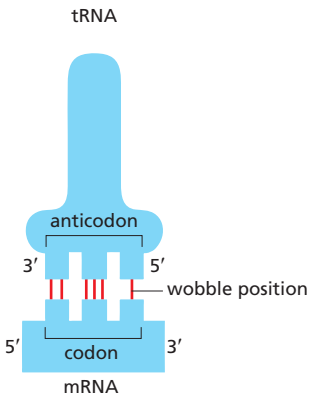


Figure 6–50 A tRNA molecule. A tRNA specific for the amino acid phenylalanine (Phe) is depicted in various ways. (A) The cloverleaf structure showing the complementary base-pairing *(red lines)* that creates the double-helical regions of the molecule. The anticodon is the sequence of three nucleotides 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 6–41) and D (dihydrouridine—see Figure 6–53) 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. (D) The tRNA icon we use in this book. (E) The linear nucleotide sequence of the molecule, color-coded to match (A), (B), and (C).

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 6–50**). 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 (see Figure 6–50B and C).

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 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 saw above that the genetic code is redundant; that is, several different codons can specify a single amino acid. 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 6–51**). This wobble base-pairing explains why so many of the alternative codons for an amino acid differ only in their third nucleotide (see Figure 6–48). In bacteria, wobble base-pairings make it possible to fit the 20 amino acids to their 61 codons with as



bacteria

wobble codon base	possible anticodon bases
U	A, G, or I
C	G or I
A	U or I
G	C or U

eukaryotes

wobble codon base	possible anticodon bases
U	A, G, or I
C	G or I
A	U
G	C

Figure 6–51 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 eukaryotes. The inosine in tRNAs is formed from the deamination of adenosine (see Figure 6–53), 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. Codon–anticodon base-pairing is more stringent at positions 1 and 2 of the codon, where only conventional base pairs are permitted. The differences in wobble base-pairing interactions between bacteria and eukaryotes presumably result from subtle structural differences between bacterial and eukaryotic 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.)

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, and among them 48 different anticodons are represented.

tRNAs Are Covalently Modified Before They Exit from the Nucleus

Like most other eukaryotic RNAs, tRNAs are covalently modified before they are allowed to exit from the nucleus. Eukaryotic tRNAs are synthesized by RNA polymerase III. Both bacterial and eukaryotic 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 eukaryotes) 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 6–52**). 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 serve 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 6–53**. 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 6–51). Others affect the accuracy with which the tRNA is attached to the correct amino acid.

Figure 6–52 Structure of a tRNA-splicing endonuclease docked to a precursor tRNA. The endonuclease (a four-subunit enzyme) removes the tRNA intron *(dark blue, bottom).* 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; PDB code: 2A9L.)

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 6–54** and **Figure 6–55**). 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

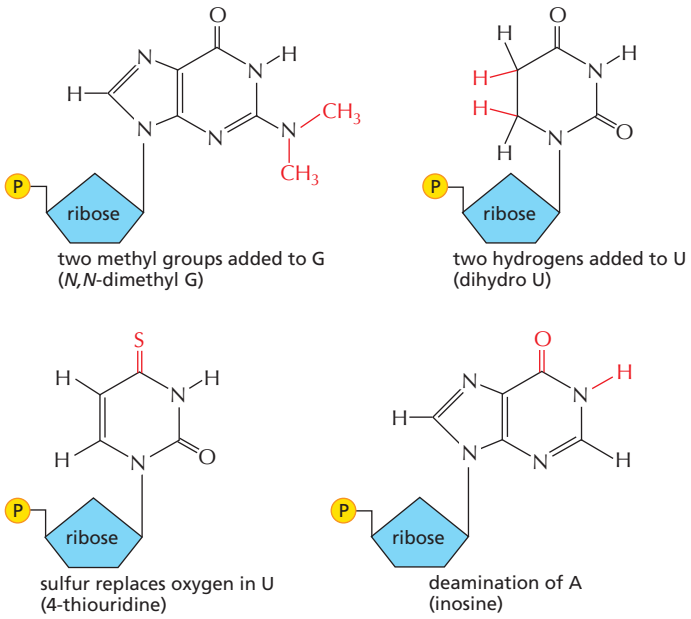


Figure 6–53 A few of the unusual nucleotides found in tRNA molecules. These nucleotides are produced by covalent modification of a normal nucleotide after it has been incorporated into an RNA chain. Two other types of modified nucleotides are shown in Figure 6–41. In most tRNA molecules, about 10% of the nucleotides are modified (see Figure 6–50). As shown in Figure 6–51, inosine is sometimes present at the wobble position in the tRNA anticodon.

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.

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. 64–65), 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 6–56). This was established by an experiment in

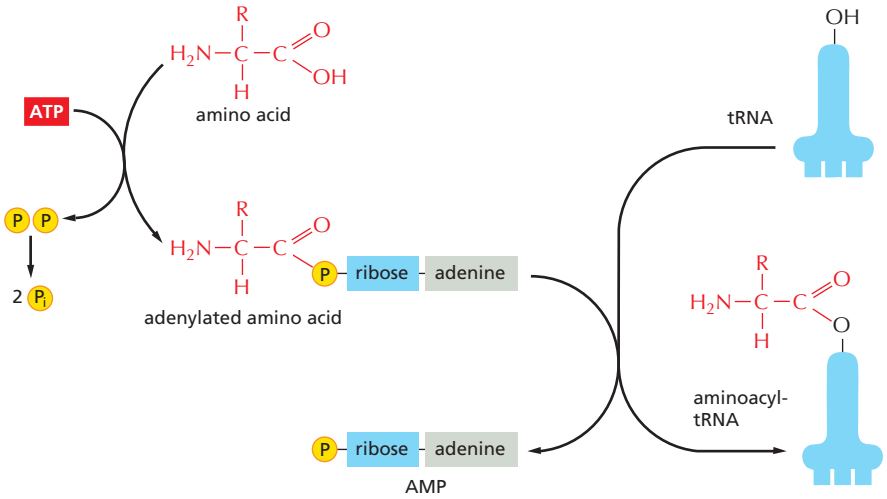


Figure 6–54 Amino acid activation by synthetase enzymes. An amino acid is activated for protein synthesis by an aminoacyl-tRNA synthetase enzyme in two steps. As indicated, the energy of ATP hydrolysis is used to attach each amino acid to its tRNA molecule in a high-energy linkage. The amino acid is first activated through the linkage of its carboxyl group directly to AMP, 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.

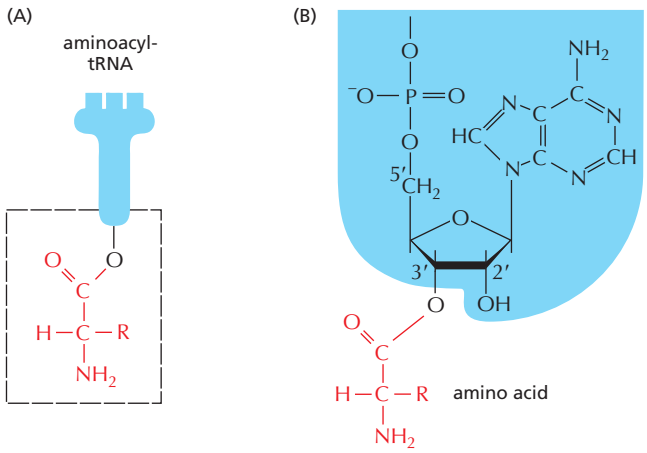


Figure 6–55 The structure of the aminoacyl-tRNA linkage. The carboxyl end of the amino acid forms an ester bond to ribose. Because the hydrolysis of this ester bond is associated with a large favorable change in free energy, an amino acid held in this way is said to be activated. (A) Schematic drawing of the structure. The amino acid is linked to the nucleotide at the 3′ end of the tRNA (see Figure 6–50). (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 6–54, the "R group" indicates the side chain of the amino acid.

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 did establish 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 an aminoacyl-tRNA synthetase links the correct amino acid to each tRNA. Most synthetase enzymes select the correct amino acid by a two-step mechanism. The correct amino acid has the highest affinity for the active-site pocket of its synthetase and is therefore favored over the other 19; in particular, amino acids larger than the correct one are excluded from the active site. However, accurate discrimination between two similar amino acids, such as isoleucine and valine (which differ by only a methyl

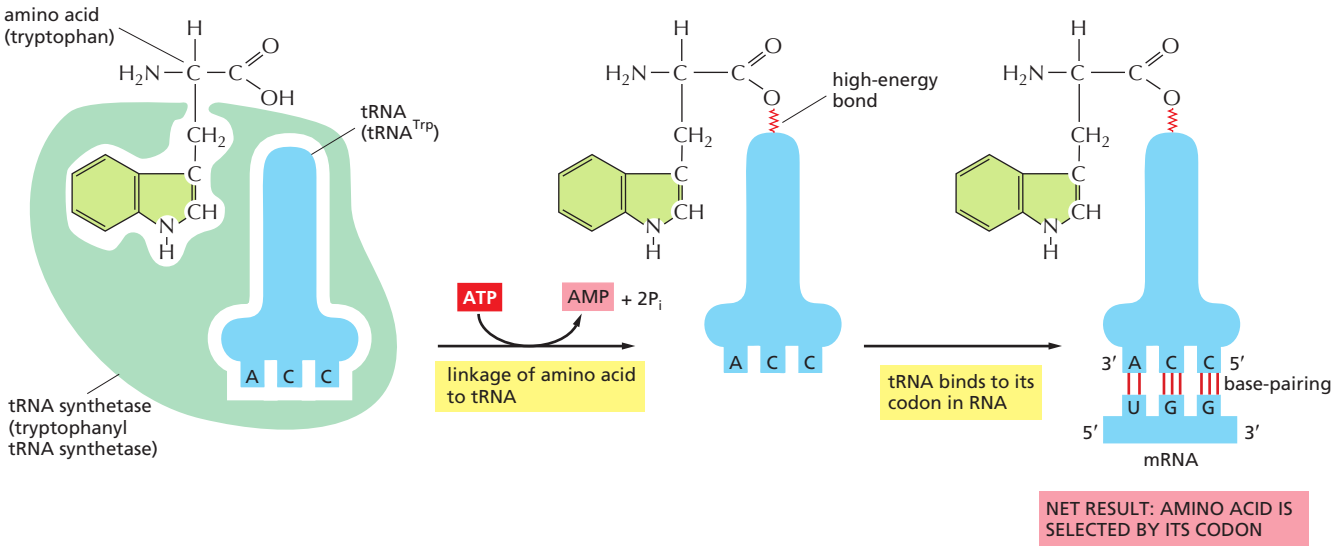


Figure 6–56 The genetic code is translated by means of two adaptors that act one after another. The first adaptor is the aminoacyl-tRNA synthetase, which couples a particular amino acid to its corresponding tRNA; the second adaptor is the tRNA molecule itself, whose *anticodon* forms base pairs with the appropriate *codon* on the mRNA. An error in either step would cause the wrong amino acid to be incorporated into a protein chain (**Movie 6.6**). In the sequence of events shown, the amino acid tryptophan (Trp) is selected by the codon UGG on the mRNA.

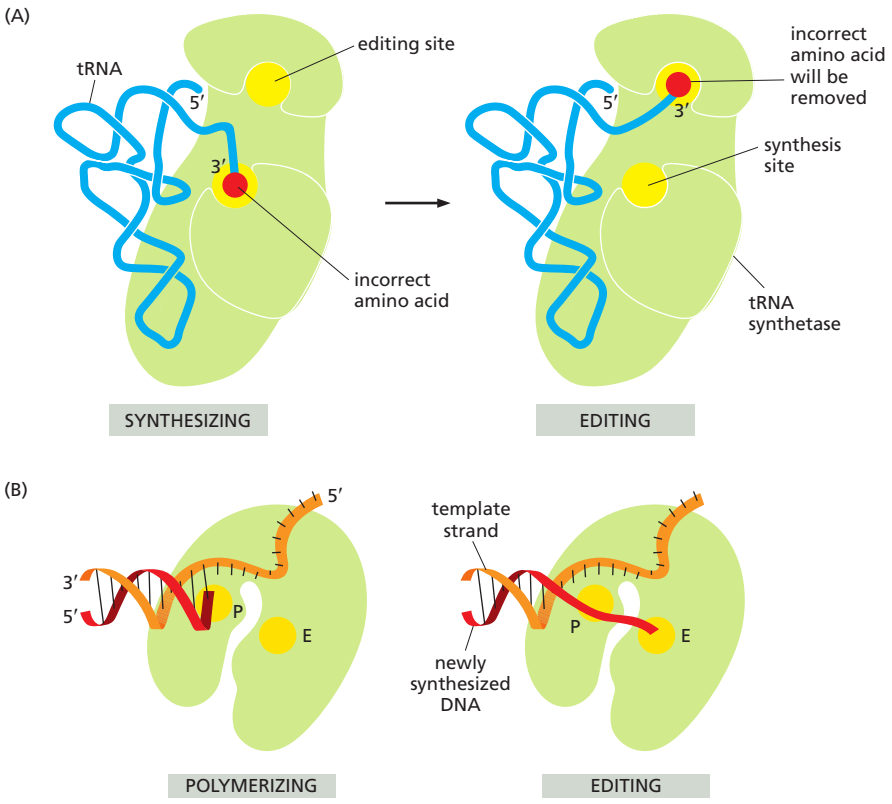


Figure 6–57 Hydrolytic editing. (A) Aminoacyl tRNA synthetases correct their own coupling errors through hydrolytic editing of incorrectly attached amino acids. As described in the text, the correct amino acid is rejected by the editing site. (B) The error-correction process performed by DNA polymerase has similarities; however, it differs because the removal process depends strongly on a mispairing with the template (see Figure 5–8). (P, polymerization site; E, editing site.)

group), is very difficult to achieve in a single step. A second discrimination step occurs after the amino acid has been covalently linked to AMP (see Figure 6–54): when tRNA binds, the synthetase tries to force the adenylated amino acid into a second editing pocket in the enzyme. The precise dimensions of this pocket exclude the correct amino acid, while allowing access by closely related amino acids. In the editing pocket, an amino acid is removed from the AMP (or from the tRNA itself if the aminoacyl-tRNA bond has already formed) by hydrolysis. This hydrolytic editing, which is analogous to the exonucleolytic proofreading by DNA polymerases, increases the overall accuracy of tRNA charging to approximately one mistake in 40,000 couplings (**Figure 6–57**).

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 6–58**). 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 amino acid-accepting arm (acceptor stem) is the key recognition determinant. In most cases, however, the synthetase "reads" the nucleotides at several different positions on the tRNA.

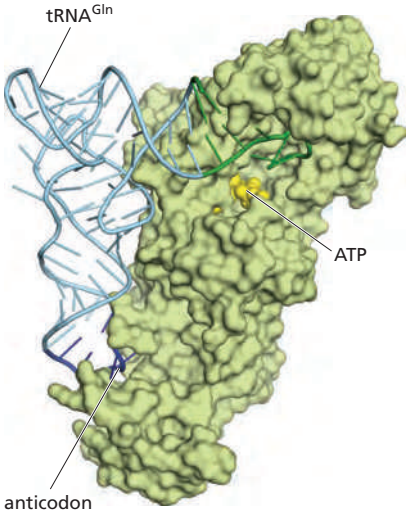


Figure 6–58 The recognition of a tRNA molecule by its aminoacyl-tRNA synthetase. For this tRNA (tRNA^{Gln}), specific nucleotides in both the anticodon *(dark blue)* and the amino acid-accepting arm *(green)* allow the correct tRNA to be recognized by the synthetase enzyme *(yellow-green)*. A bound ATP molecule is *yellow*. (Courtesy of Tom Steitz; PDB code: 1QRS.)

Amino Acids Are Added to the C-terminal End of a Growing Polypeptide Chain

Having seen that each amino acid is first coupled to specific tRNA molecules, we now turn to the mechanism that joins these 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

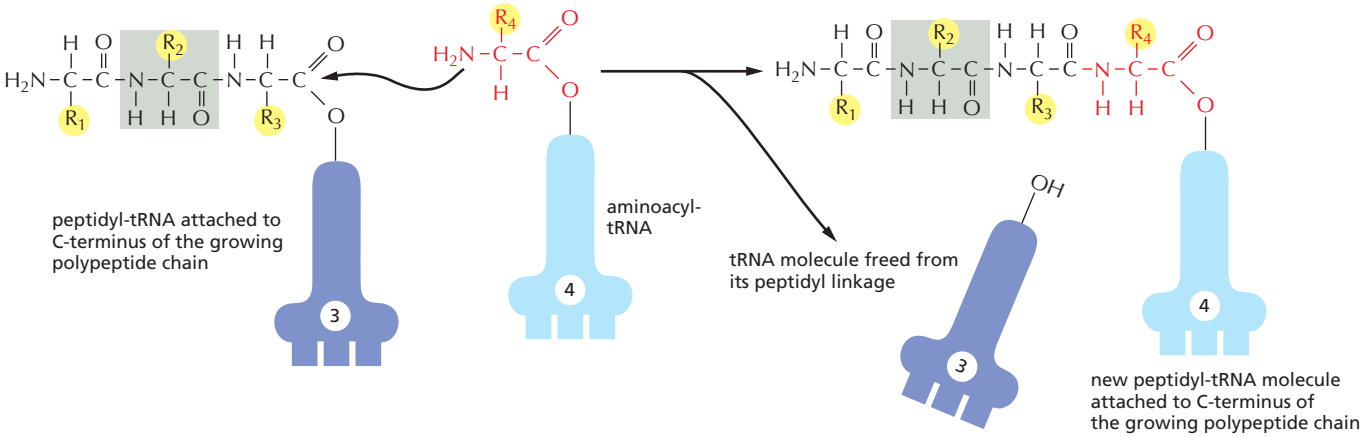


Figure 6–59 The incorporation of an amino acid into a protein. A polypeptide chain grows by the stepwise addition of amino acids to its C-terminal end. The formation of each peptide bond is energetically favorable because the growing C-terminus has been activated by the covalent attachment of a tRNA 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.

addition disrupts this high-energy covalent linkage, but immediately replaces it with an identical linkage on the most recently added amino acid (**Figure 6–59**). 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 2–44.

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). A typical eukaryotic cell contains millions of ribosomes in its cytoplasm (**Figure 6–60**). The large and small ribosome subunits are assembled at the nucleolus, where newly transcribed and modified rRNAs associate with the ribosomal proteins that have been transported into the nucleus after their synthesis in the cytoplasm. These two ribosomal subunits are then exported to the cytoplasm, where they join together to synthesize proteins.

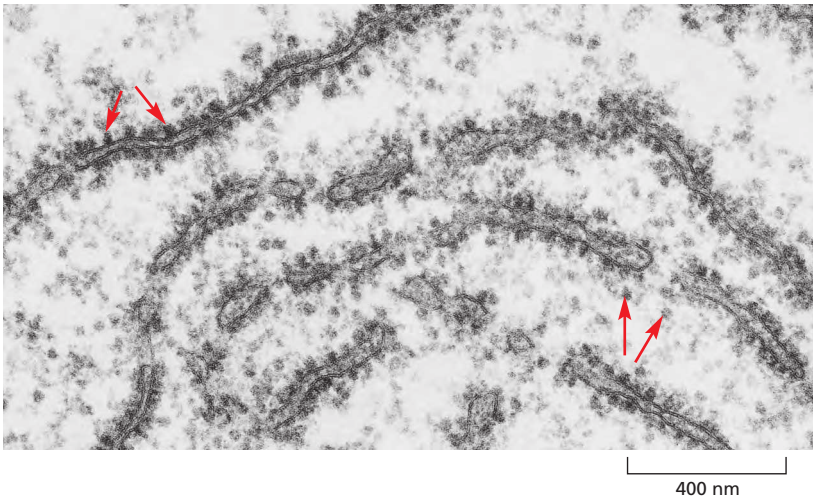


Figure 6–60 Ribosomes in the cytoplasm of a eukaryotic cell. This electron micrograph shows a thin section of a small region of cytoplasm. The ribosomes appear as black dots *(red arrows)*. Some are free in the cytosol; others are attached to membranes of the endoplasmic reticulum. (Courtesy of Daniel S. Friend.)

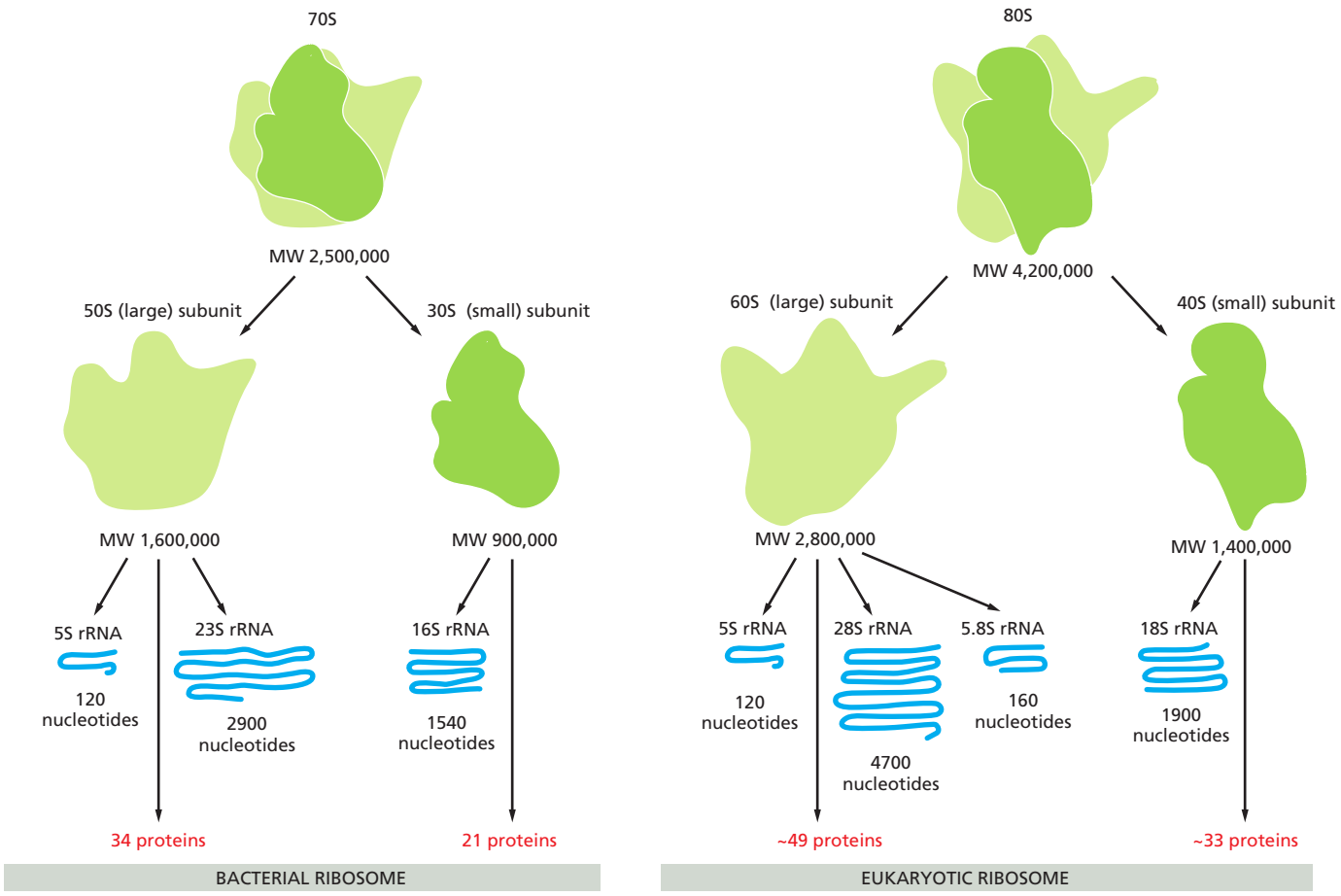


Figure 6–61 A comparison of bacterial and eukaryotic ribosomes. Despite differences in the number and size of their rRNA and protein components, both bacterial and eukaryotic ribosomes have nearly the same structure and they function similarly. Although the 18S and 28S rRNAs of the eukaryotic ribosome 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 the rRNA largely unchanged.

Eukaryotic and bacterial ribosomes have similar structures and functions, being composed of one large and one small subunit that fit together to form a complete ribosome with a mass of several million daltons (**Figure 6–61**). The small subunit provides the framework on which the tRNAs are accurately matched to the codons of the mRNA, while the large subunit catalyzes the formation of the peptide bonds that link the amino acids together into a polypeptide chain (see Figure 6–58).

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, three nucleotides at a time. 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 growing end of the 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 eukaryotic ribosome adds 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.

To 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 6–62**). A tRNA molecule is held tightly at the A and P sites only if its anticodon

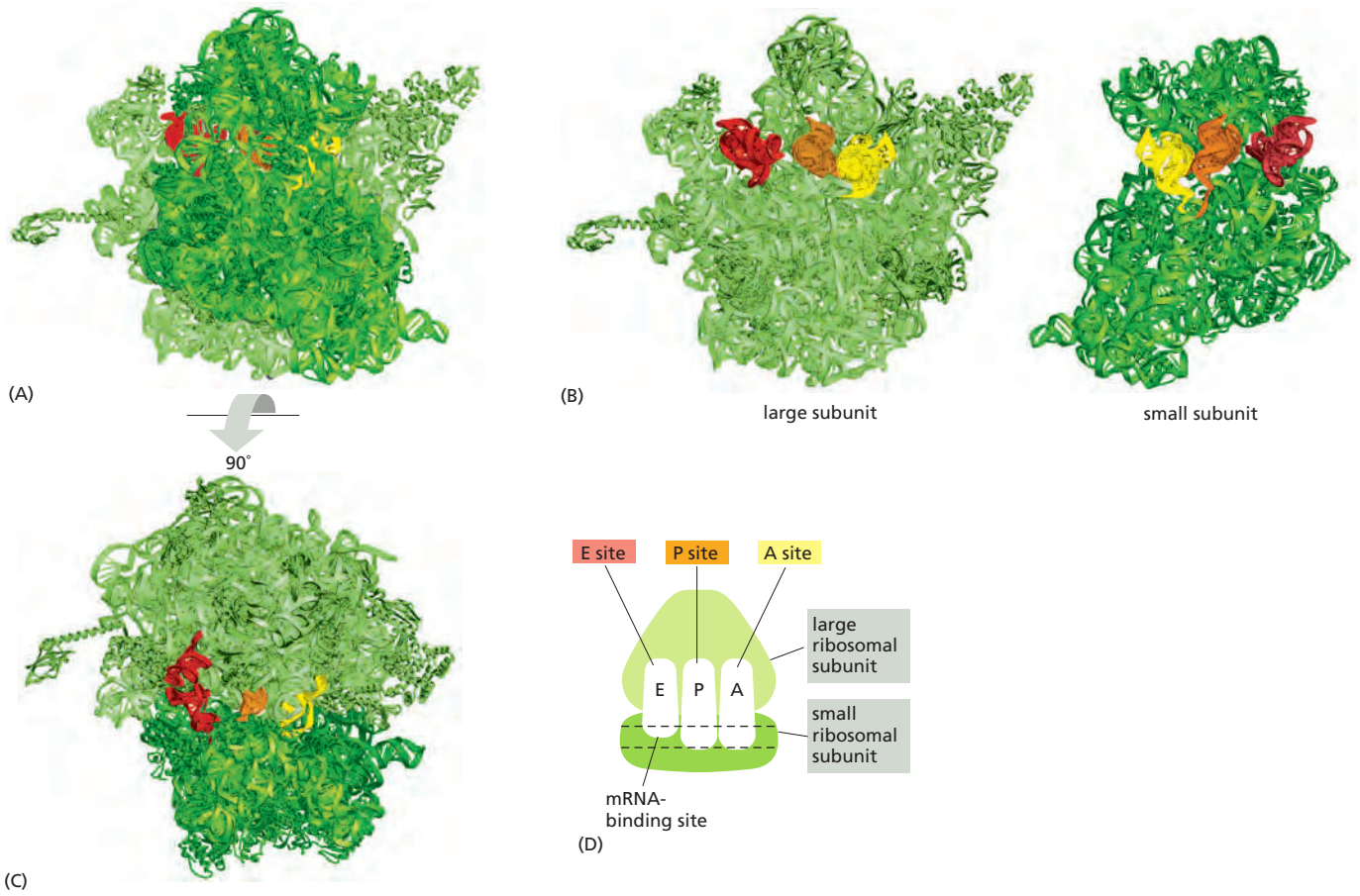


Figure 6–62 The RNA-binding sites in the ribosome. Each ribosome has one binding site for mRNA and three binding sites for tRNA: the A, P, and E sites (short for aminoacyl-tRNA, peptidyl-tRNA, and exit, respectively). (A) A bacterial ribosome viewed with the small subunit in the front *(dark green)* and the large subunit in the back *(light green).* Both the rRNAs and the ribosomal proteins are illustrated. 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 6–64). (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 90° 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.)

forms base pairs with a complementary codon (allowing for wobble) on the mRNA molecule that is threaded through the ribosome (**Figure 6–63**). 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 binding (step 1), peptide bond formation (step 2), large subunit translocation (step 3), and small subunit translocation (step 4). 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 6–64** illustrates this four-step process, beginning at a point at which three amino acids have already been linked together and there is a tRNA molecule in the P site on the ribosome, covalently joined to the C-terminal end of the short 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)

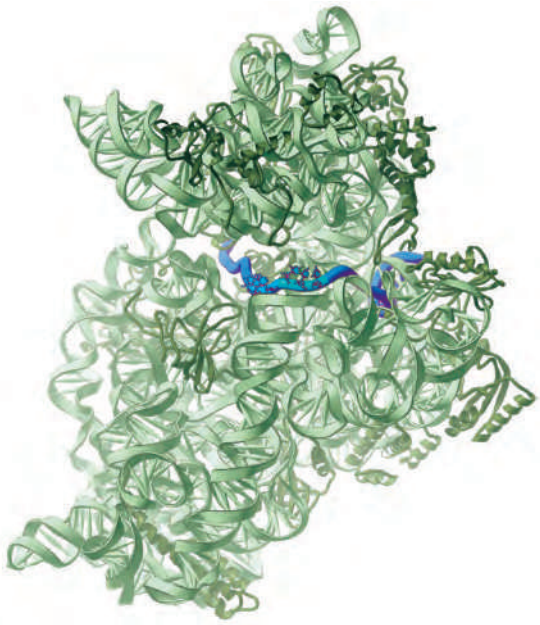


Figure 6–63 The path of mRNA *(blue)* through the small ribosomal subunit. The orientation is the same as that in the right-hand panel of Figure 6–62B. (Courtesy of Harry F. Noller, based on data in G.Z. Yusupova et al., *Cell* 106:233–241, 2001. With permission from Elsevier.)

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, ejecting the spent tRNA from the E site and 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.

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.

Elongation Factors Drive Translation Forward and Improve Its Accuracy

The basic cycle of polypeptide elongation shown in outline in Figure 6–64 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 eukaryotes. Under some conditions *in vitro,* ribosomes can be forced to synthesize proteins

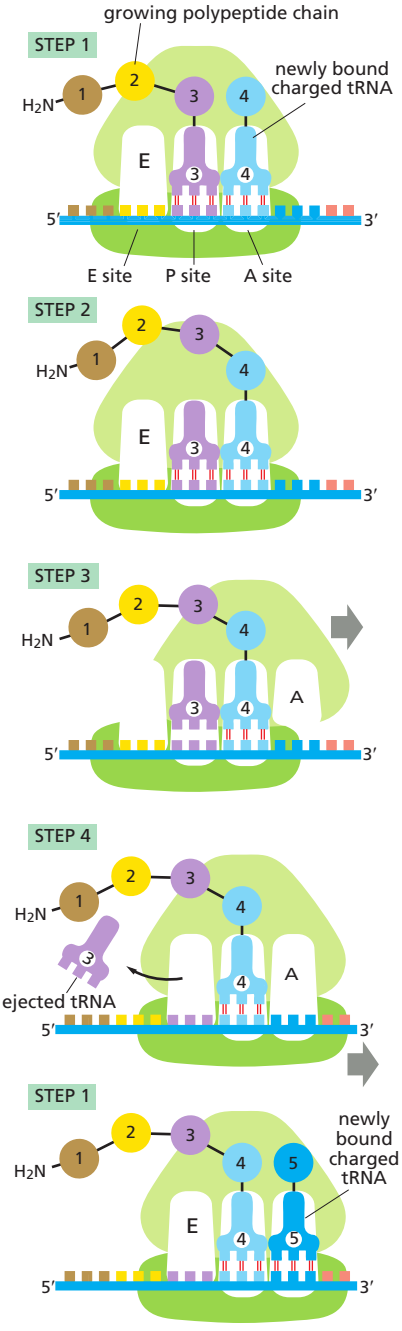


Figure 6–64 Translating an mRNA molecule. Each amino acid added to the growing end of a polypeptide chain is selected by complementary base-pairing between the anticodon on its attached tRNA molecule and the next codon on the mRNA chain. Because only one of the many types of tRNA molecules in a cell 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. 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 (Movie 6.7 and Movie 6.8).

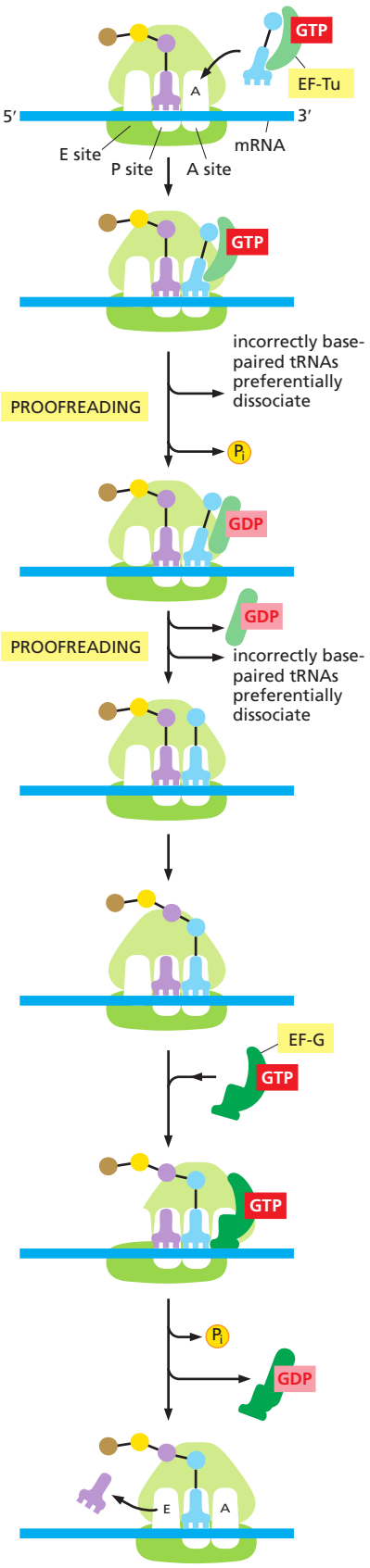


Figure 6–65 Detailed view of the translation cycle. The outline of translation presented in Figure 6–64 has been expanded to show the roles of the two elongation factors EF-Tu and EF-G, which drive translation in the forward direction. As explained in the text, EF-Tu provides opportunities for proofreading of the codon–anticodon match. In this way, incorrectly paired tRNAs are selectively rejected, and the accuracy of translation is improved. The binding of a molecule of EF-G to the ribosome and the subsequent hydrolysis of GTP lead to a rearrangement of the ribosome structure, moving the mRNA being decoded exactly three nucleotides through it (Movie 6.9).

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. The cycles of elongation factor association, GTP hydrolysis, and dissociation also ensure that all such changes occur in the "forward" direction, helping translation to proceed efficiently (Figure 6–65).

In addition to moving translation forward, EF-Tu increases its accuracy. As we discussed in Chapter 3, EF-Tu can simultaneously bind GTP and aminoacyl-tRNAs (see Figures 3–72 and 3–73), and it is in this form that the initial codon–anticodon interaction occurs in the A site of the ribosome. Because of the free-energy change associated with base-pair formation, a correct codon–anticodon match will bind more tightly than an incorrect interaction. However, this difference in affinity is relatively modest and cannot by itself account for the high accuracy of translation.

To increase the accuracy of this binding reaction, the ribosome and EF-Tu work together in the following ways. First, the 16s rRNA in the small subunit of the ribosome assesses the "correctness" of the codon–anticodon match by folding around it and probing its molecular details (Figure 6–66). When a correct match is found, the rRNA closes tightly around the codon–anticodon pair, causing a conformational change in the ribosome that triggers GTP hydrolysis by EF-Tu. Only when GTP is hydrolyzed does EF-Tu release its grip on the aminoacyl-tRNA and allow it to be used in protein synthesis. Incorrect codon–anticodon matches do not readily trigger this conformational change, and these errant tRNAs mostly fall off the ribosome before they can be used in protein synthesis. Proofreading, however, does not end here.

After GTP is hydrolyzed and EF-Tu dissociates from the ribosome, there is a second opportunity for the ribosome to prevent an incorrect amino acid from being added to the growing chain. 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 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. The two proofreading steps, acting in series, are largely responsible for the 99.99% accuracy of the ribosome in translating RNA into protein.

Even if the wrong amino acid slips through the proofreading steps just described and is incorporated onto the growing polypeptide chain, there is still one more opportunity for the ribosome to detect the error and provide a solution, albeit one that is not, strictly speaking, proofreading. An incorrect codon-anticodon interaction in the P site of the ribosome (which would occur *after* the misincorporation) causes an increased rate of misreading in the A site. Successive rounds of amino acid misincorporation eventually lead to premature termination of the protein by *release factors*, which are described below. Normally, these release factors act when translation of a protein is complete; here, they act early. Although this mechanism does not correct the original error, it releases the flawed protein for degradation, ensuring that no additional peptide synthesis is wasted on it.

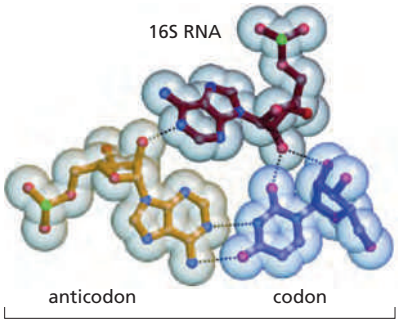


Figure 6–66 Recognition of correct codon–anticodon matches by the small-subunit rRNA of the ribosome. Shown here is the interaction between a nucleotide of the small-subunit rRNA and the first nucleotide pair of a correctly paired codon–anticodon. Similar interactions form between other nucleotides of the rRNA and the second and third positions of codon–anticodon pair. The small-subunit rRNA can form this network of hydrogen bonds only when an anticodon is correctly matched to a codon. 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.)

Many Biological Processes Overcome the Inherent Limitations of Complementary Base-Pairing

We have seen in this and the previous chapter that DNA replication, repair, transcription, and translation all rely on complementary base-pairing—G with C, and A with T (or U). However, if only the difference in hydrogen bonding is considered, a correct versus incorrect match should differ in affinity only by a factor of 10- to 100-fold. These processes have an accuracy much higher than can be accounted for by this difference. Although the mechanisms used to "squeeze out" additional specificity from complementary base-pairing differ from one process to the next, two principles exemplified by the ribosome appear to be general.

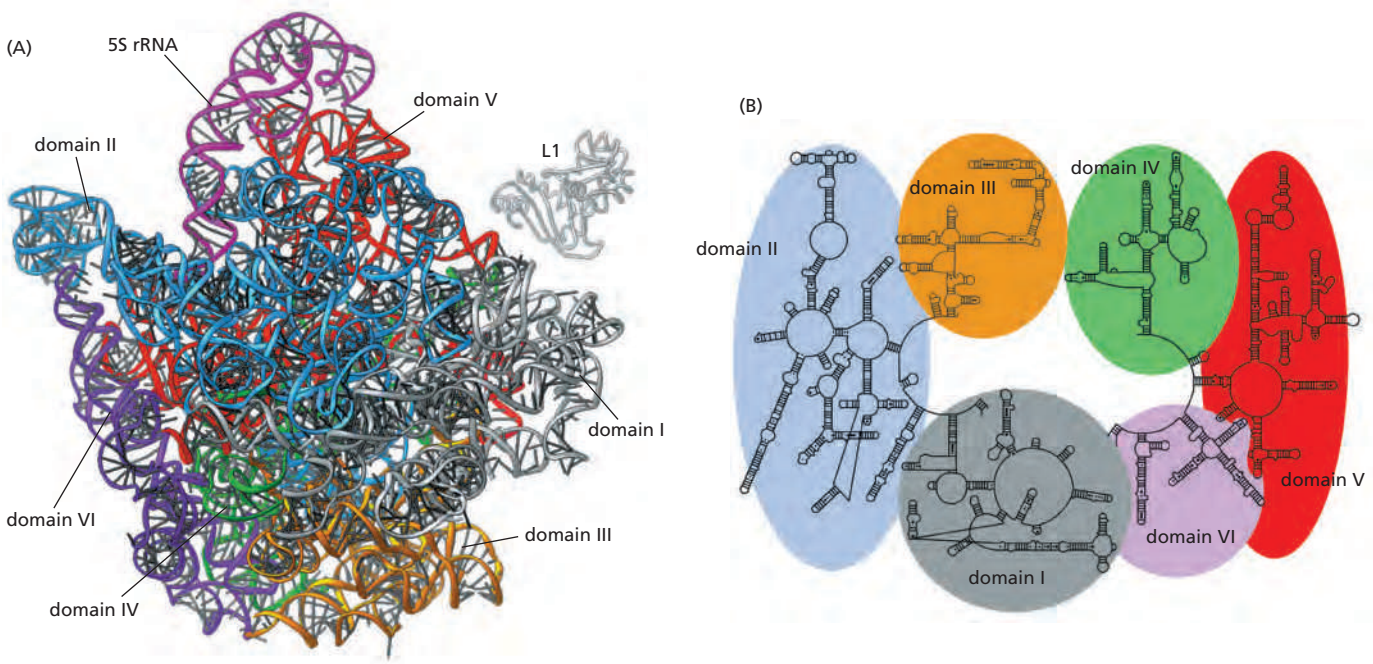
The first is **induced fit**. We have seen that, before an amino acid is added to a growing polypeptide chain, the ribosome folds around the codon–anticodon interaction, and only when the match is correct is this folding completed and the reaction allowed to proceed. Thus, the codon–anticodon interaction is thereby checked twice—once by the initial complementary base-pairing and a second time by the folding of the ribosome, which depends on the correctness of the match. This same principle of induced fit is seen in transcription by RNA polymerase; here, an incoming nucleoside triphosphate initially forms a base pair with the template; at this point the enzyme folds around the base pair (thereby assessing its correctness) and, in doing so, creates the active site of the enzyme. The enzyme then covalently adds the nucleotide to the growing chain. Because their geometry is "wrong," incorrect base pairs block this induced fit, and they are therefore likely to dissociate before being incorporated into the growing chain.

A second principle used to increase the specificity of complementary base-pairing is called **kinetic proofreading**. We have seen that after the initial codon-anticodon pairing and conformational change of the ribosome, GTP is hydrolyzed. This creates an irreversible step and starts the clock on a time delay during which the aminoacyl-tRNA moves into the proper position for catalysis. During this delay, those incorrect codon–anticodon pairs that have somehow slipped through the induced-fit scrutiny have a higher likelihood of dissociating than correct pairs. There are two reasons for this: (1) the interaction of the wrong tRNA with the codon is weaker, and (2) the delay is longer for incorrect than correct matches.

In its most general form, kinetic proofreading refers to a time delay that begins with an irreversible step such as ATP or GTP hydrolysis, during which an incorrect substrate is more likely to dissociate than a correct one. In this case, kinetic proofreading thus increases the specificity of complementary base-pairing above what is possible from simple thermodynamic associations alone. The increase in specificity produced by kinetic proofreading comes at an energetic cost in the form of ATP or GTP hydrolysis. Kinetic proofreading is believed to operate in many biological processes, but its role is understood particularly well for translation.

Accuracy in Translation Requires an 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 6–54), and two more drive steps in the cycle of reactions occurring on the ribosome during protein synthesis itself (see Figure 6–65). In addition, extra energy is consumed each time that an incorrect amino acid linkage is hydrolyzed by a tRNA synthetase (see Figure 6–57) and each time that an incorrect tRNA enters the ribosome, triggers GTP hydrolysis, and is rejected (see Figure 6–65). To be effective, any proofreading mechanism 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 at first seem.

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 6–67**).

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 6–68**). Some of these proteins send out extended regions of polypeptide chain that penetrate short distances into holes in the RNA core (**Figure 6–69**). The main role of the ribosomal proteins seems to be to stabilize the

Figure 6–67 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 rRNA 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.)

RNA core, while permitting the changes in rRNA conformation that are necessary for this RNA to catalyze efficient protein synthesis. The proteins 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. 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. An additional surprise came from the discovery that the tRNA in the P site contributes an important OH group to the active site and participates directly in the catalysis. This mechanism may ensure that catalysis occurs only when the P site tRNA is properly positioned in the ribosome.

RNA molecules that possess catalytic activity are known as **ribozymes**. We saw earlier in this chapter that some ribozymes function in self-splicing reactions. In the final section of this chapter, we consider what the ability of RNA molecules to function as catalysts 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.

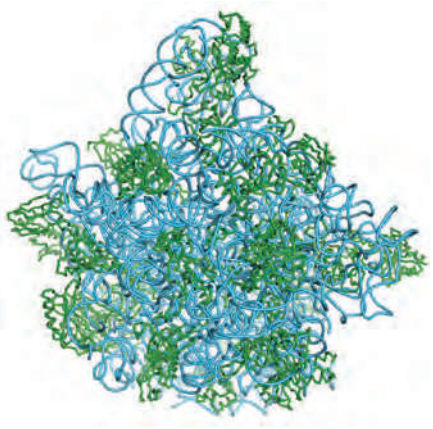


Figure 6–68 Location of the protein components of the bacterial large ribosomal subunit. The rRNAs (5S and 23S) are shown in *blue* and the proteins of the large subunit in *green*. This view is toward the outside of the ribosome; the interface with the small subunit is on the opposite face. (PDB code: 1FFK.)

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 to produce a protein. The rate of this step is thus one determinant of the rate at which any particular protein will be synthesized. We shall see in Chapter 7 how regulation of this step occurs.

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 is specially recognized by initiation factors because it has a nucleotide sequence distinct from that of the tRNA that normally carries methionine.

In eukaryotes, the initiator tRNA–methionine complex (Met–tRNAi) is first loaded into the small ribosomal subunit along with additional proteins called **eukaryotic initiation factors**, or **eIFs**. 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 unlike other tRNAs it binds directly to the P site (**Figure 6–70**). Next, the small ribosomal subunit binds to the 5′ end of an mRNA molecule, which is recognized by virtue of its 5′ cap that has previously bound two initiation factors, eIF4E and eIF4G (see Figure 6–38). 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

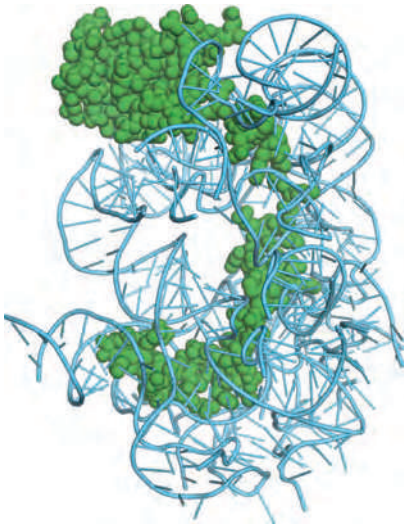


Figure 6–69 Structure of the L15 protein in the large subunit of the bacterial ribosome. The globular domain of the protein lies on the surface of the ribosome and an extended region penetrates deeply into the RNA core of the ribosome. The L15 protein is shown in *green* and a portion of the ribosomal RNA core is shown in *blue*. (From D. Klein, P.B. Moore and T.A. Steitz, *J. Mol. Biol.* 340:141–177, 2004. With permission from Academic Press. PDB code: 1S72.)

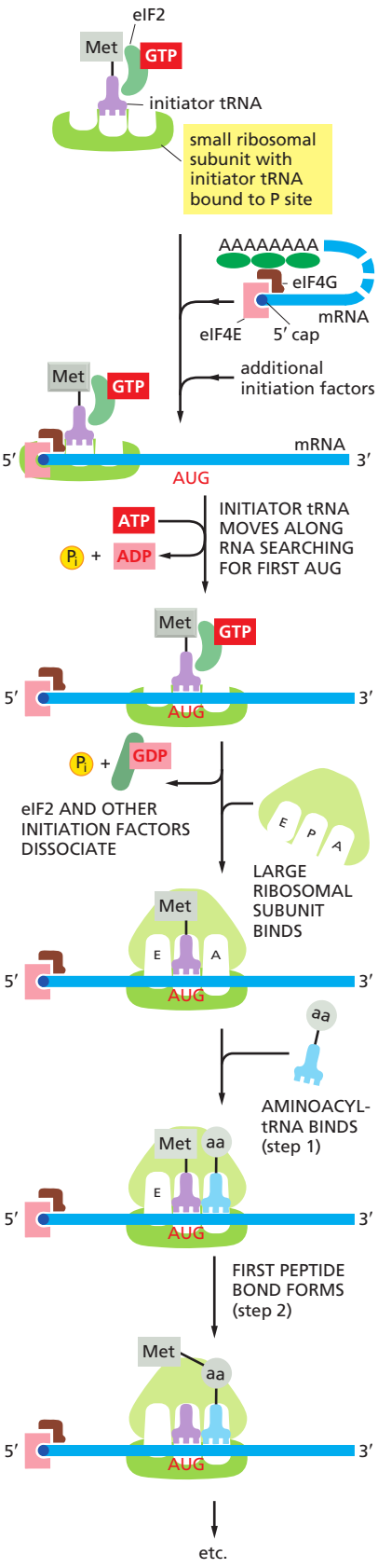


Figure 6–70 The initiation of protein synthesis in eukaryotes. Only three of the many translation initiation factors required for this process are shown. Efficient translation initiation also requires the poly-A tail of the mRNA bound by poly-A-binding proteins, which, in turn, interact with eIF4G (see Figure 6–38). In this way, the translation apparatus ascertains that both ends of the mRNA are intact before initiating protein synthesis. 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. In the last two steps shown in the figure, the ribosome has begun the standard elongation cycle, depicted in Figure 6–64.

helicases facilitate this movement. 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 remains at the P site, leaving the A site vacant. Protein synthesis is therefore ready to begin (see Figure 6–70).

The nucleotides immediately surrounding the start site in eukaryotic 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. This mechanism 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 eukaryotic ribosome, a bacterial ribosome can 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 6–71**). In contrast, a eukaryotic mRNA generally encodes only a single protein, or more accurately, a single set of closely related proteins.

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 6–48). 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 6–72**). 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 its bound mRNA molecule and separates into the large and small subunits. These subunits can then assemble on this or another mRNA molecule to begin a new round of protein synthesis.

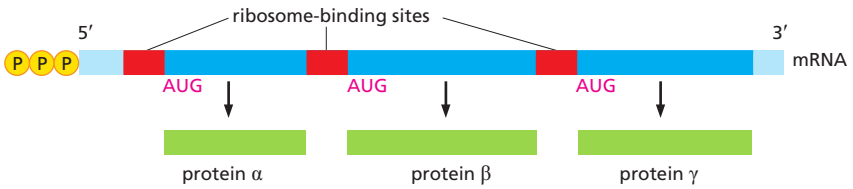


Figure 6–71 Structure of a typical bacterial mRNA molecule. Unlike eukaryotic ribosomes, which typically require a capped 5′ end on the mRNA, prokaryotic ribosomes initiate translation at ribosome-binding sites (Shine–Dalgarno sequences), which can be located anywhere along an mRNA molecule. This property of their ribosomes permits bacteria to synthesize more than one type of protein from a single mRNA molecule.

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. 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. 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 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 6–73**). These multiple initiations allow the cell to make many more protein molecules in a given time than would be possible if each protein had to be completed before the next could start.

Both bacteria and eukaryotes use polysomes, and both employ additional strategies to speed up the overall rate of protein synthesis. 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 eukaryotes, as we have seen, the 5′ and 3′ ends of the mRNA interact (see Figure 6–73A); 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 6–48) applies to all three major branches of life, providing important evidence for the common 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

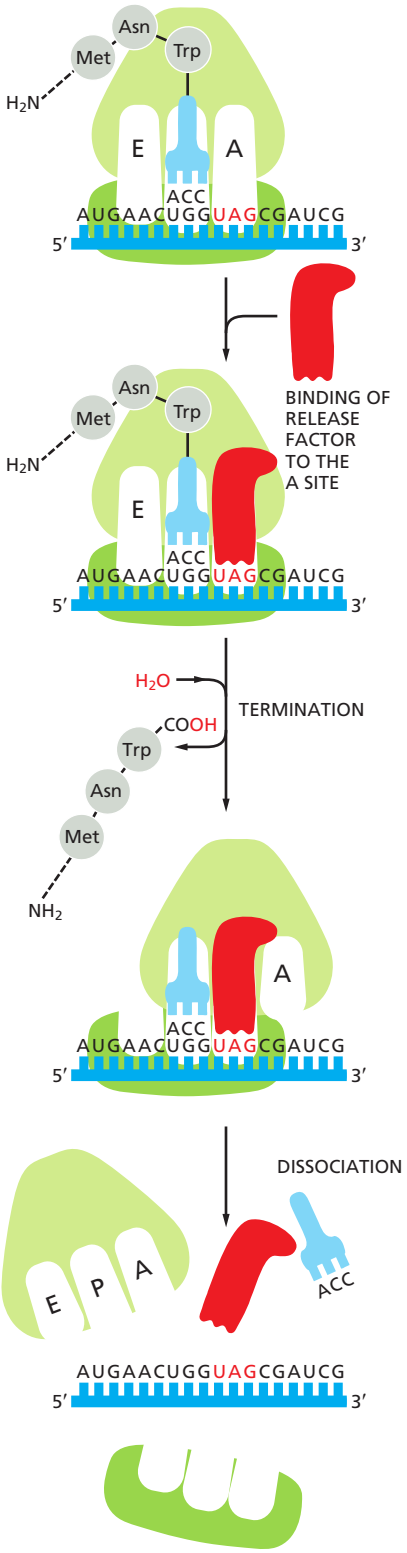


Figure 6–72 The final phase of protein synthesis. The binding of a release factor to an A site bearing a stop codon terminates translation. The completed polypeptide is released and, in a series of reactions that requires additional proteins and GTP hydrolysis (not shown), the ribosome dissociates into its two separate subunits.

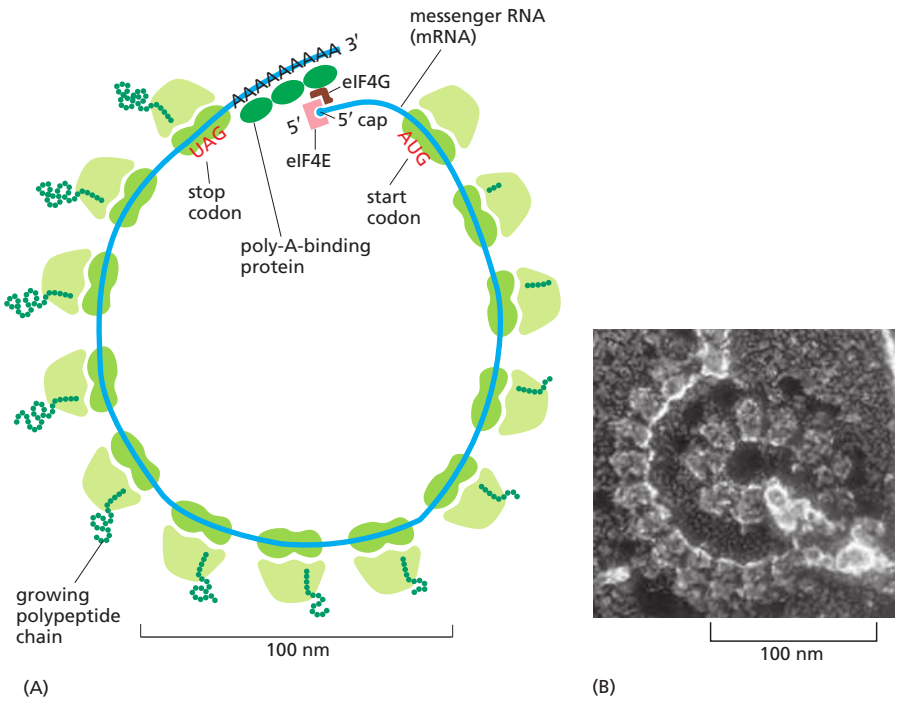


Figure 6–73 A polyribosome. (A) Schematic drawing showing how a series of ribosomes can simultaneously translate the same eukaryotic mRNA molecule. (B) Electron micrograph of a polyribosome from a eukaryotic cell (Movie 6.10). (B, courtesy of John Heuser.)

cytosol of the cell it is translated as isoleucine (see Table 14–3, p. 805). 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 eukaryotes 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 nearby nucleotide sequence in the mRNA that triggers this recoding event (Figure 6–74).

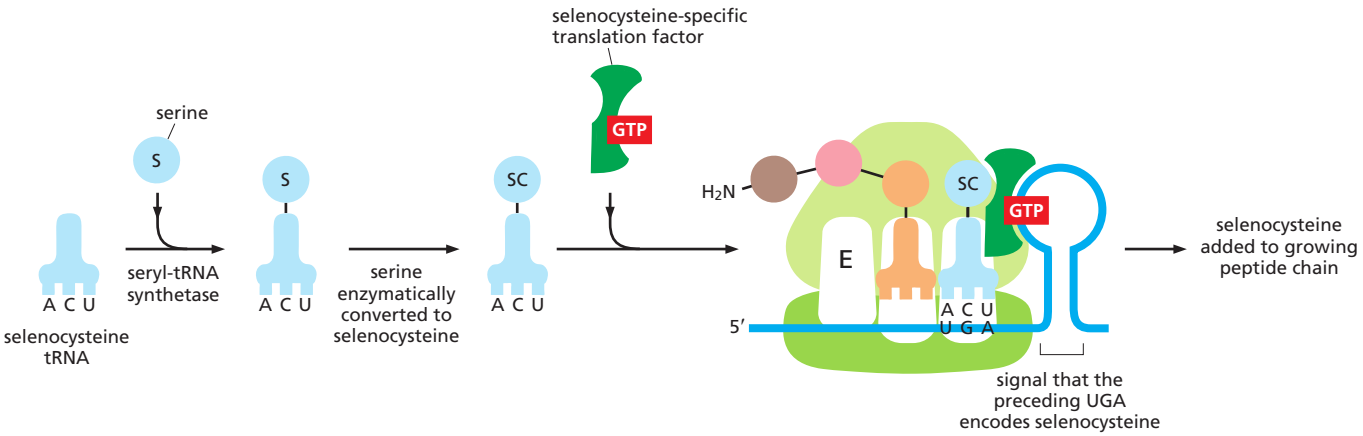


Figure 6–74 Incorporation of selenocysteine into a growing polypeptide chain. A specialized tRNA is charged with serine by the normal seryl-tRNA synthetase, and the serine is subsequently converted enzymatically to selenocysteine. A specific RNA structure in the mRNA (a stem and loop structure with a 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. After the addition of selenocysteine, translation continues until a conventional stop codon is encountered.

Inhibitors of Prokaryotic 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 have resulted in fungi producing potent bacterial inhibitors. Some of these drugs exploit the structural and functional differences between bacterial and eukaryotic 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; others block specific parts of the ribosome such as the exit channel (**Figure 6–75**). **Table 6–4** lists some common antibiotics of this kind along with several other inhibitors of protein synthesis, some of which act on eukaryotic 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 eukaryotic cell, for example, chloramphenicol inhibits protein synthesis on ribosomes only in mitochondria (and in chloroplasts in plants), presumably reflecting the prokaryotic 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 prokaryotes and eukaryotes.

Quality Control Mechanisms Act to Prevent Translation of Damaged mRNAs

In eukaryotes, mRNA production involves both transcription and a series of elaborate RNA processing steps; as we have seen, these take place in the nucleus, segregated from ribosomes, and only when the processing is complete are the mRNAs transported to the cytosol to be translated (see Figure 6–38). However, this scheme is not foolproof, and some incorrectly processed mRNAs are inadvertently sent to the cytosol. In addition, mRNAs that were flawless when they left the nucleus can become broken or otherwise damaged in the cytosol. The danger of

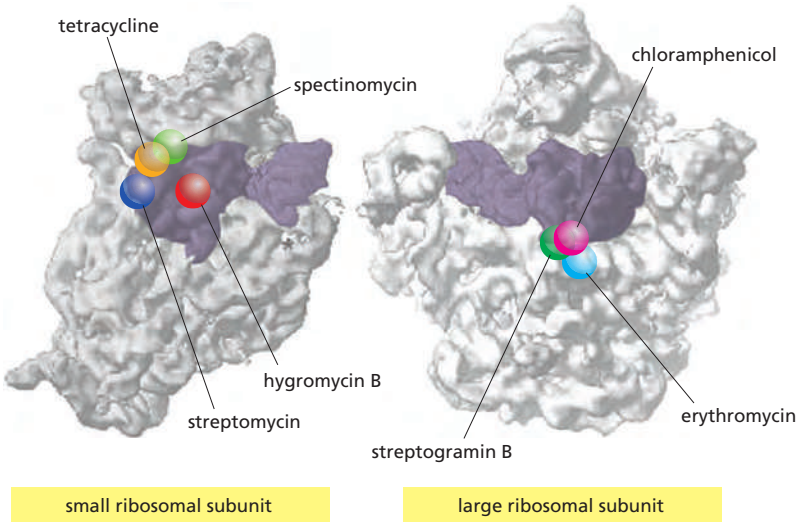


Figure 6–75 Binding sites for antibiotics on the bacterial ribosome. The small *(left)* and large *(right)* subunits of the ribosome are arranged as though the ribosome has been opened like a book. Antibiotic binding sites are marked with colored spheres, and the bound tRNA molecules are shown in *purple* (see Figure 6–62). 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 the 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 6–64)
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 eukaryotes	
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 eukaryotes but not bacteria	
Cycloheximide	Blocks the translocation reaction on ribosomes (step 3 in Figure 6–64)
Anisomycin	Blocks the peptidyl transferase reaction on ribosomes (step 2 in Figure 6–64)
α-Amanitin	Blocks mRNA synthesis by binding preferentially to RNA polymerase II

The ribosomes of eukaryotic 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.

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, for example, the 5′ cap and the poly-A tail are both recognized by the translation-initiation machinery before translation begins (see Figure 6–70).

The most powerful mRNA surveillance system, called **nonsense-mediated mRNA decay**, eliminates defective mRNAs before they move away from the nucleus. 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. This situation is likely to arise in an mRNA molecule that has been improperly spliced, because 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 6–31B).

The nonsense-mediated mRNA decay mechanism begins as an mRNA molecule is being transported from the nucleus to the cytosol. As its 5′ end emerges from a nuclear pore, the mRNA is met by a ribosome, which begins to translate it. As translation proceeds, the exon junction complexes (EJCs) that are bound to the mRNA at each splice site are displaced by the moving ribosome. The normal stop codon will lie within the last exon, so by the time the ribosome reaches it and stalls, no more EJCs will be bound to the mRNA. In this case, the mRNA "passes inspection" and is released to the cytosol where it can be translated in earnest (**Figure 6–76**). However, if the ribosome reaches a stop codon earlier, when EJCs remain bound, the 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 eukaryotic 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 24–28) often generate

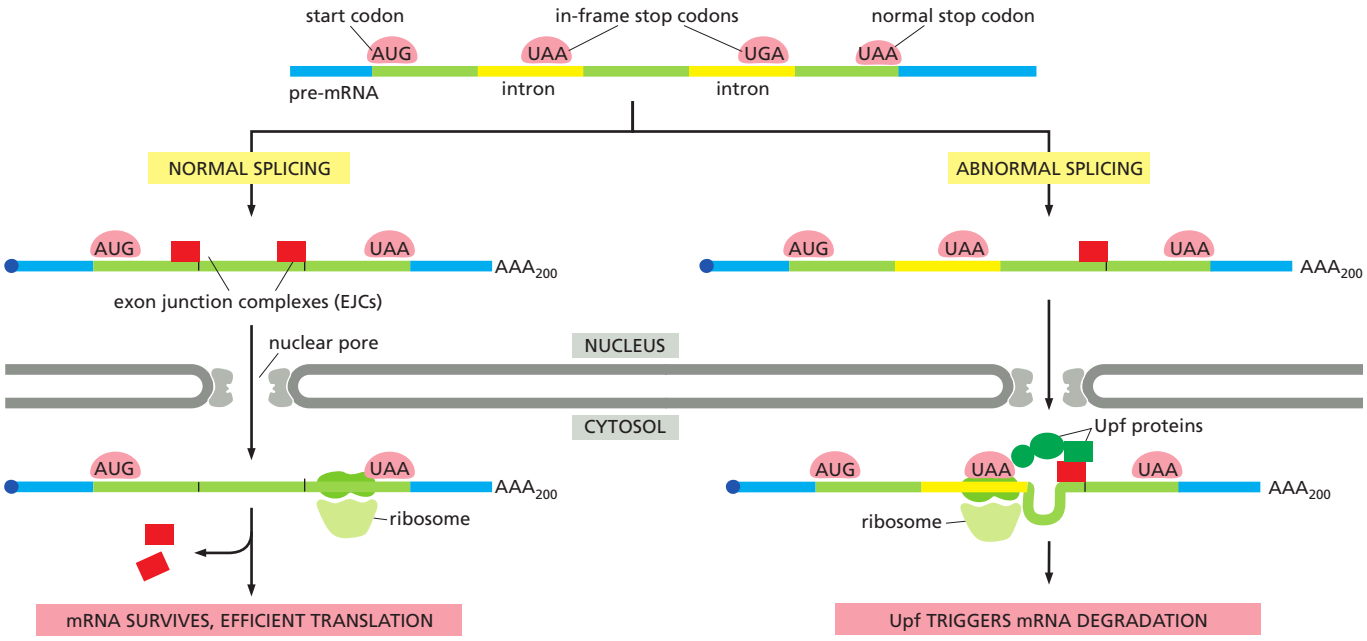


Figure 6–76 Nonsense-mediated mRNA decay. As shown on the right, the failure to correctly splice a pre-mRNA often introduces a premature stop codon into the reading frame for the protein. These abnormal mRNAs are destroyed by the nonsense-mediated decay mechanism. To activate this mechanism, an mRNA 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 EJC 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 this mechanism ensures that nonsense-mediated decay is triggered only when the premature stop codon is 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.)

premature termination codons. The surveillance system degrades the mRNAs produced from such rearranged genes, thereby avoiding the potential toxic effects of truncated proteins.

The nonsense-mediated surveillance pathway also 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.

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 6–77**).

The information needed for all of the steps listed above is ultimately contained in the sequence of amino acids that the ribosome produces when it translates an mRNA molecule into a polypeptide chain. As discussed in Chapter 3, when a

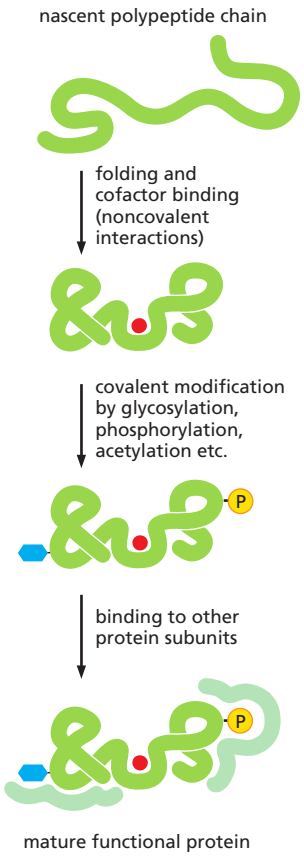


Figure 6–77 Steps in the creation of a functional protein. As indicated, translation of an mRNA sequence into an amino acid sequence on the ribosome is not the end of the process of forming a protein. To function, the completed polypeptide chain must fold correctly into its three-dimensional conformation, 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, over 200 different types of covalent modifications are known (see pp. 165–166).

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 pp. 114–115).

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 chain emerges from 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 6–78**). For some 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 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 6–79**).

Molecular Chaperones Help Guide the Folding of Most Proteins

Most proteins probably do not fold correctly during their synthesis and require a special class of proteins called **molecular chaperones** to do so. Molecular chaperones are useful for cells because there are many different folding paths available to an unfolded or partially folded protein. Without chaperones, some of these pathways would not lead to the correctly folded (and most stable) form because the protein would become "kinetically trapped" in structures that are off-pathway. Some of these off-pathway configurations would aggregate and be left as irreversible dead ends of nonfunctional (and potentially dangerous) structures.

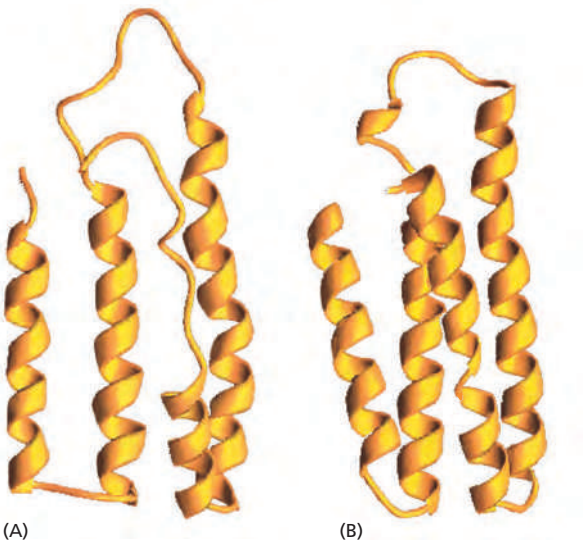


Figure 6–78 The structure of a molten globule. (A) A molten globule form of cytochrome b_{562} is more open and less highly ordered than the final folded form of the protein, shown in (B). Note that the molten globule contains most of the secondary structure of the final form, although the ends of the α helices are unraveled 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.)

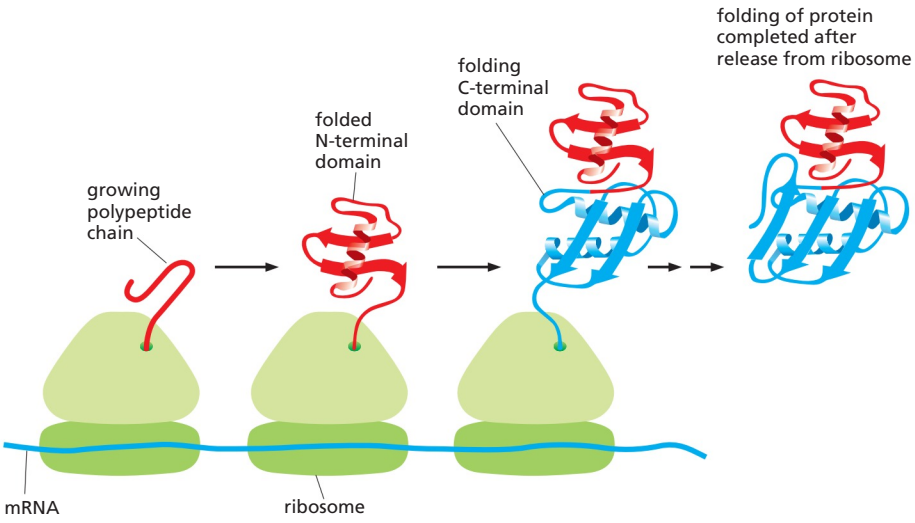


Figure 6–79 Co-translational protein folding. A growing polypeptide chain is shown acquiring its secondary and tertiary structure as it emerges from a ribosome. The N-terminal domain folds first, while the C-terminal domain is still being synthesized. This protein has not achieved its final conformation at the time it is released from the ribosome. (Modified from A.N. Fedorov and T.O. Baldwin, *J. Biol. Chem*. 272:32715–32718, 1997.)

Molecular chaperones specifically recognize incorrect, off-pathway configurations by their exposure of hydrophobic surfaces, which in correctly folded proteins are typically buried in the interior. The binding of these exposed hydrophobic surfaces to each other is what causes off-pathway conformations to irreversibly aggregate. We saw in Chapter 3 that in some cases of inherited human diseases, aggregates do form and can cause severe symptoms and even death. Chaperones prevent this from happening in normal proteins by binding to the exposed hydrophobic surfaces using hydrophobic surfaces of their own. As we shall see shortly, there are several types of chaperones; once bound to an incorrectly folded protein, they ultimately release it in a way that gives the protein another chance to fold correctly.

Cells Utilize Several Types of Chaperones

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 molecular chaperones, including the hsp60 and hsp70 proteins. Different members of these families 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.

The hsp60 and hsp70 proteins each work with their own small set of associated proteins when they help other proteins to fold. These 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 (often before the protein leaves the ribosome), with each monomer of hsp70 binding to a string

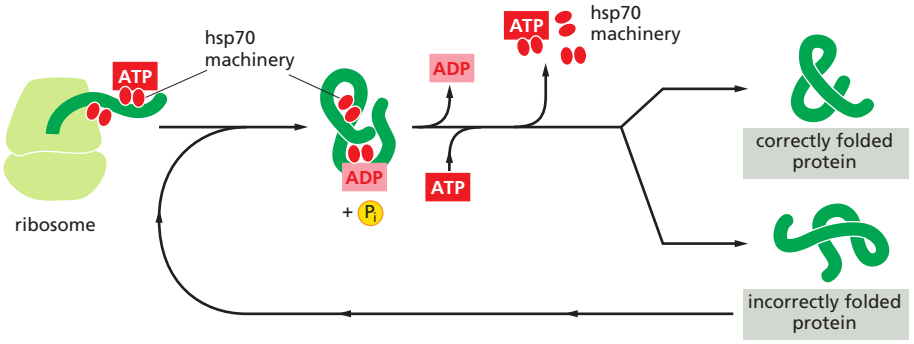
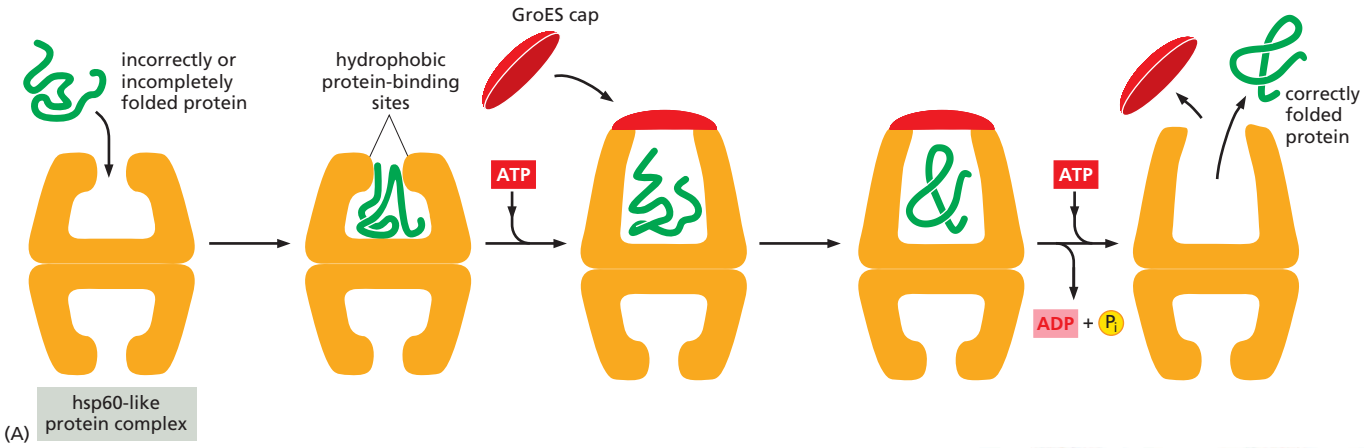


Figure 6–80 The hsp70 family of molecular chaperones. These proteins act early, recognizing a small stretch of hydrophobic amino acids on a protein's surface. Aided by a set of smaller hsp40 proteins (not shown), ATP-bound hsp70 molecules grasp their target protein and then hydrolyze ATP to ADP, undergoing 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. Repeated cycles of hsp binding and release help the target protein to refold.

of about four or five hydrophobic amino acids (**Figure 6–80**). On binding ATP, hsp70 releases the protein into solution allowing it a chance to re-fold. 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" for the folding process (**Figure 6–81**).

To enter a chamber, a substrate protein is first captured via the hydrophobic entrance to the chamber. The protein is then released into the interior of the chamber, which is lined with hydrophilic surfaces, and the chamber is sealed with a lid, a step requiring ATP. Here, the substrate is allowed to fold into its final conformation in isolation, where there are no other proteins with which to aggregate. When ATP is hydrolyzed, the lid pops off, and the substrate protein, whether folded or not, is released from the chamber.

The chaperones shown in Figures 6–80 and 6–81 often need many cycles of ATP hydrolysis to fold a single polypeptide chain correctly. This energy is used to perform mechanical movements of the hsp60 and hsp70 "machines," converting them from binding forms to releasing forms. 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 the folding of a protein in an orderly way.



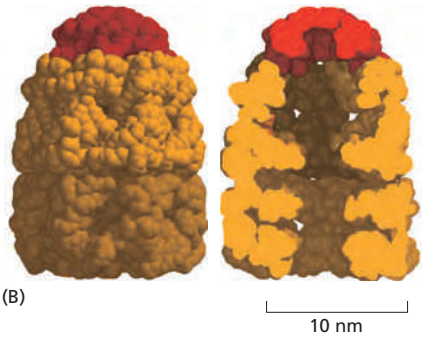


Figure 6–81 The structure and function of the hsp60 family of molecular chaperones. (A) A misfolded protein is initially captured by hydrophobic interactions with the exposed surface of the opening. The initial binding often helps to unfold a misfolded protein. The subsequent binding of ATP and a cap releases the substrate protein into an enclosed space, where it has a new opportunity to fold. After about 10 seconds, ATP hydrolysis occurs, weakening the binding of the cap. Subsequent binding of additional ATP molecules ejects the cap, and the protein is released. As indicated, only half of the symmetric barrel operates on a client protein at any one time. 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. (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.)

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 forms. 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. We have seen that the cell distinguishes misfolded proteins, which require additional rounds of ATP-catalyzed refolding, from those with correct structures through the recognition of hydrophobic surfaces.

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.

Proteins that rapidly fold correctly on their own do not display such patterns and generally bypass the chaperones. For the others, the chaperones can carry out "protein repair" by giving them additional chances to fold while, at the same time, preventing their aggregation.

Figure 6–82 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, an additional mechanism is called into play that completely destroys the protein by proteolysis. This 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 recognize 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 may be 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, and 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.

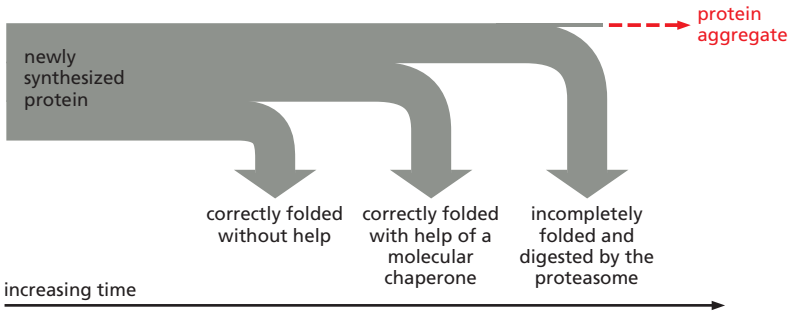


Figure 6–82 The processes that monitor protein quality following protein synthesis. A newly synthesized protein sometimes folds correctly and assembles on its own with its partner proteins, in which case the quality control mechanisms leave it alone. Incompletely folded proteins are helped to properly fold 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 substrate 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 hydrophobic 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.

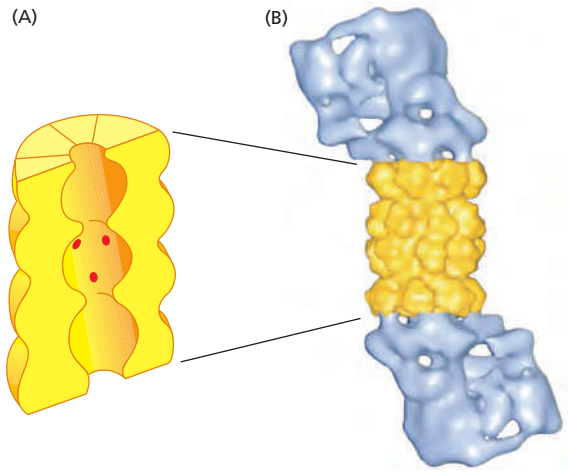


Figure 6–83 The proteasome. (A) A cut-away view of the structure of the central 20S cylinder, as determined by x-ray crystallography, with the active sites of the proteases indicated by *red dots.* (B) The entire proteasome, in which the central cylinder *(yellow)* is supplemented by a 19S cap *(blue)* at each end. The 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 6–85) 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.)

Present in many copies dispersed throughout the cytosol and the nucleus, the proteasome also destroys aberrant proteins that have entered the endoplasmic reticulum (ER). In the latter case, an ER-based surveillance system detects proteins that have failed either to fold or to be assembled properly after they enter the ER, and *retrotranslocates* them back to the cytosol for degradation by the proteasome (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 stack of four heptameric rings (**Figure 6–83**). Some of the subunits are proteases whose active sites face the cylinder's inner chamber, thus preventing them from running rampant through the cell. Each end of the cylinder is normally associated with a large protein complex (the 19S cap) that contains a six-subunit protein ring through which target proteins are threaded into the proteasome core, where they are degraded (**Figure 6–84**). 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 6–85**). 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 they share mechanistic features with the ATP-dependent DNA helicases that unwind DNA (see Figure 5–14).

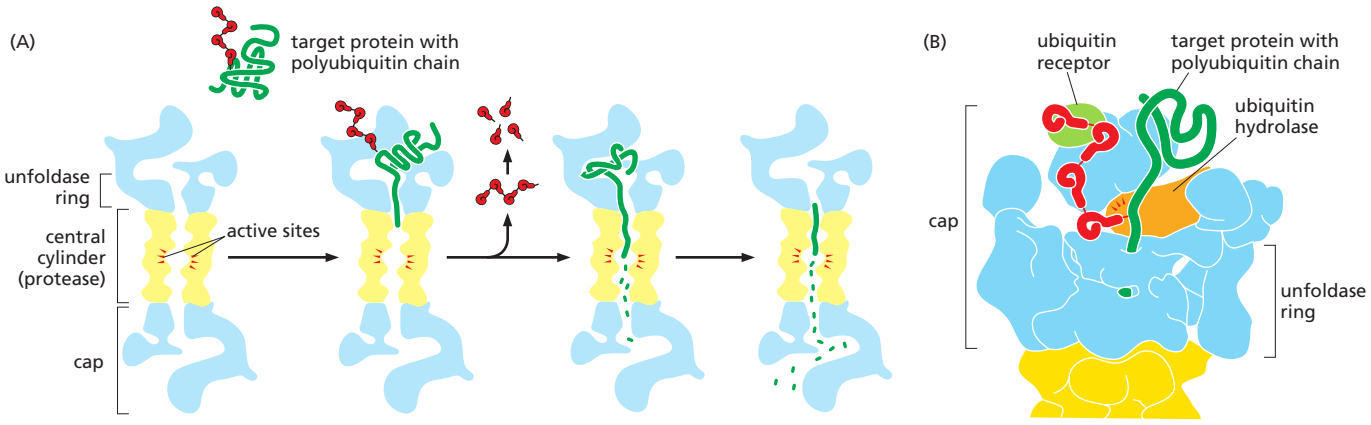
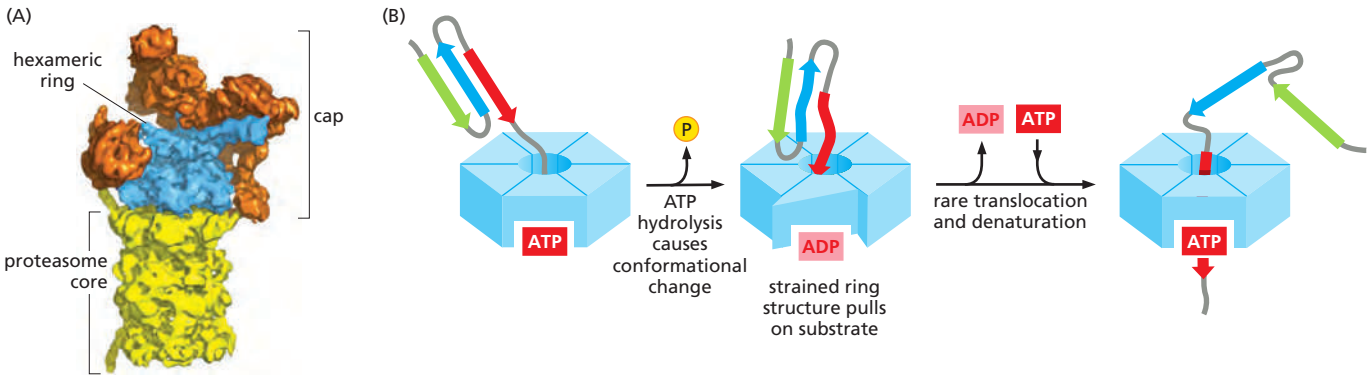


Figure 6–84 Processive protein digestion by the proteasome. (A) The proteasome cap recognizes proteins marked by a polyubiquitin chain (see Figure 3–70), and subsequently translocates them into the proteasome core, where they are digested. At an early stage, the ubiquitin is cleaved from the substrate protein and is recycled. Translocation into the core of the proteasome is mediated by a ring of ATPases that unfold the substrate protein as it is threaded through the ring and into the proteasome core. This unfoldase ring is depicted in Figure 6–85). (B) Detailed structure of the proteasome cap. The cap includes a ubiquitin receptor, which holds a ubiquitylated protein in place while it begins to be pulled into the proteasome core, and a ubiquitin hydrolase, which cleaves ubiquitin from the doomed protein. (A, from S. Prakash and A. Matouschek, *Trends Biochem. Sci.* 29:593–600, 2004. With permission from Elsevier. B, adapted from G.C. Lander et al., *Nature* 482:186–191, 2012.)



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.

One would expect that a machine as efficient as the proteasome would be tightly regulated; in particular, it must be able to distinguish abnormal proteins from those that are properly folded. The 19S cap of the proteasome acts as a gate at the entrance to the inner proteolytic core, and only those proteins marked for destruction are threaded through the cap. The destruction "mark" is the covalent attachment of the small protein ubiquitin. As we saw in Chapter 3, ubiquitin modification of proteins is used for many purposes in the cell. The particular type of ubiquitin linkage that concerns us here is a chain of ubiquitin molecules linked together at lysine 48 (see Figure 3–69); this is the distinguishing feature of the ubiquitin tag that marks a protein for destruction in the proteasome.

A special set of E3 molecules (see Figure 3–70B) is responsible for the ubiquitylation of denatured or otherwise misfolded proteins, as well as proteins containing oxidized or other abnormal amino acids. Abnormal proteins tend to display on their surface hydrophobic amino acid sequences or conformational motifs that are recognized as degradation signals by these E3 molecules; these sequences are buried and therefore inaccessible in the normal, properly folded version. However, a proteolytic pathway that recognizes and destroys abnormal proteins must be able to distinguish *completed* proteins that have "wrong" conformations from 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; in the course of carrying out its main job, the ubiquitin–proteasome system probably destroys many nascent and newly formed protein molecules, not because these proteins are abnormal as such, but because they have transiently exposed degradation signals that are buried in their mature (folded) state.

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. Indeed, every protein in the cell eventually accumulates damage and is probably degraded by the proteasome. 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.

Figure 6–85 A hexameric protein unfoldase. (A) The proteasome cap includes proteins *(orange)* that recognize and hydrolyze ubiquitin and a hexameric ring *(blue)* through which ubiquitylated proteins are threaded. The hexameric ring 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 is held in place by its ubiquitin tag. A conformational change, driven 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 partially withdraw. Very stable protein substrates may require hundreds of cycles of ATP hydrolysis and dissociation before they are successfully pulled through the AAA protein ring. Once unfolded (and de-ubiquitylated), the substrate protein moves relatively quickly through the pore by successive rounds of ATP hydrolysis. (A, adapted from G.C. Lander et al., *Nature* 482:186–191, 2012; B, adapted from R.T. Sauer et al., *Cell* 119:9–18, 2004. With permission from Elsevier.)

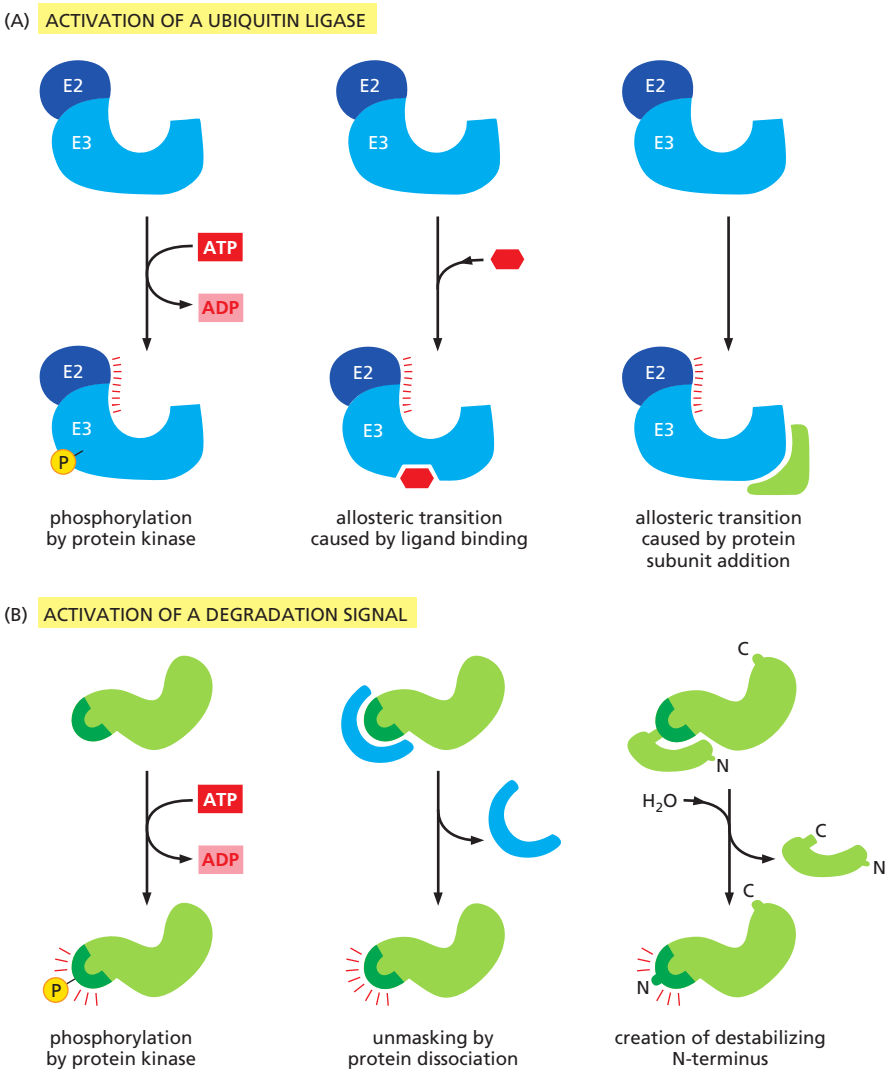


Figure 6–86 Two general ways of inducing the degradation of a specific protein. (A) Activation of a specific E3 molecule creates a new ubiquitin ligase. Eukaryotic cells have many different E3 molecules, each activated by a different signal. (B) Creation of an exposed degradation signal in the protein 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.

How is such a regulated destruction of a protein controlled? Several general mechanisms are illustrated in **Figure 6–86**. Specific examples of each mechanism are discussed in later chapters. In one general class of mechanism (Figure 6–86A), 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 17–15A).

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 (Figure 6–86B). 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 protein as a "destabilizing" N-terminal residue. This E3 protein recognizes only certain amino acids at the N-terminus of a protein; thus not all protein-cleavage events will lead to degradation of the C-terminal fragment produced.

In humans, nearly 80% of proteins are acetylated on their N-terminal residue, and we now know that this modification is recognized by a specific E3 enzyme,

which directs the ubiquitylation of the protein and sends it to the proteasome for degradation. Thus, the majority of human proteins carry their own signals for destruction. It has been proposed that when a protein is properly folded (and, before that, when it is in contact with a chaperone), this acetylated N-terminus is buried and therefore inaccessible to the E3 enzyme. According to this idea, as a protein ages and becomes damaged (or if it fails to fold correctly from the start), this destruction signal becomes exposed, and the protein is destroyed.

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 6–87**). 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. We also now know that the cell devotes enormous resources to selectively degrading proteins, particularly those that fail to fold properly or accumulate damage as they age. It is the balance between the rates of synthesis and degradation that determines the final amount of every protein in the cell.

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 6–87 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.

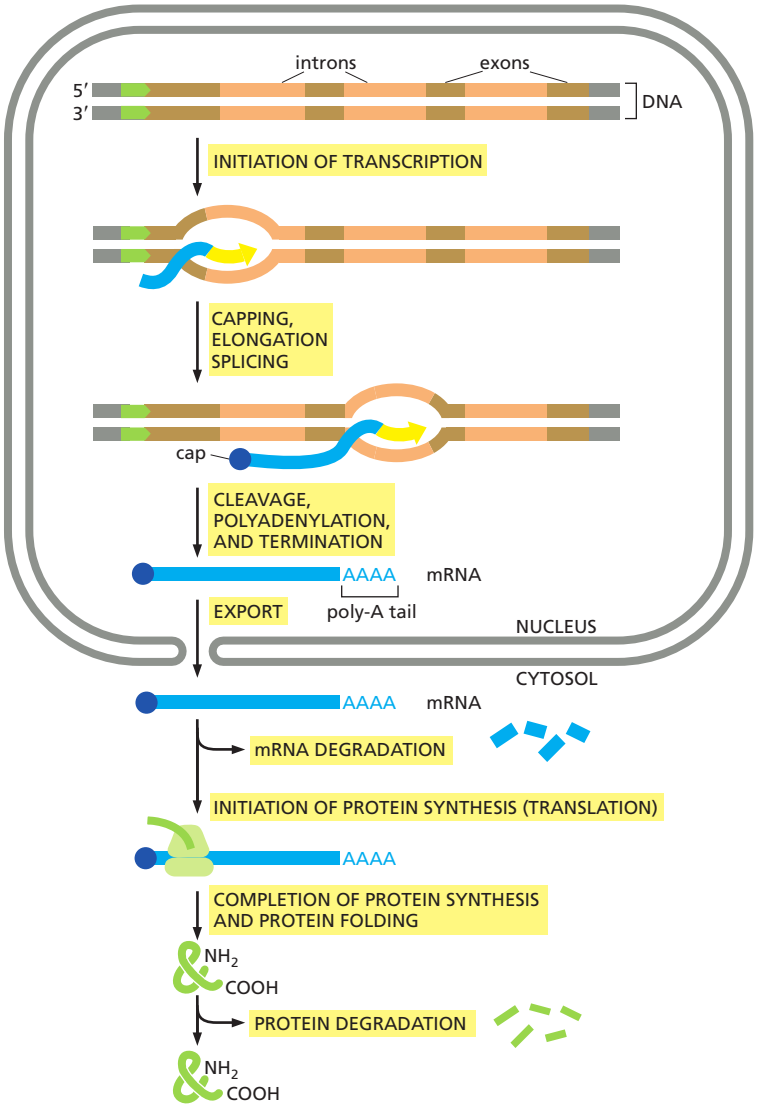


Figure 6–87 The production of a protein by a eukaryotic cell. The final level of each protein in a eukaryotic cell depends upon the efficiency of each step depicted.

Summary

The translation of the nucleotide sequence of an mRNA molecule into protein takes place in the cytosol on a large ribonucleoprotein assembly called a ribosome. Each amino acid used for protein synthesis is first attached to a tRNA molecule that recognizes, by complementary base-pair interactions, a particular set of three nucleotides (codons) in the mRNA. As an mRNA is threaded through a ribosome, its sequence of nucleotides 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 then 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 both 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.

Eukaryotic 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 unfolds the protein and threads it into 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 as well as to remove selected proteins 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 6–88**). 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 catalysts and structural components 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 from an earlier world.

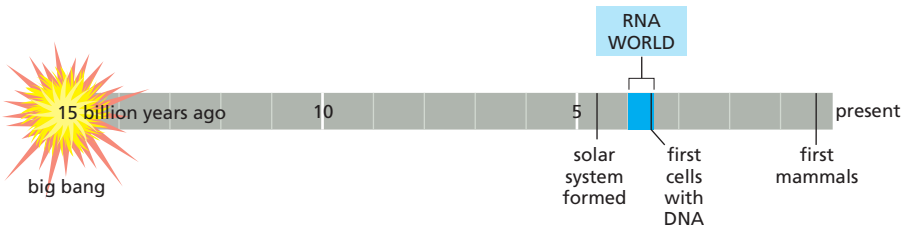
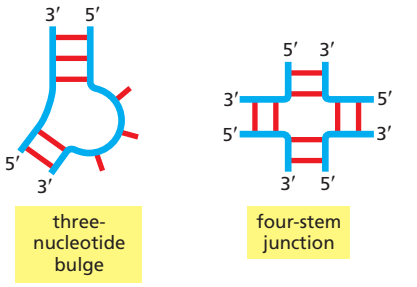


Figure 6–88 **Time line for the universe, suggesting the early existence of an RNA world of living systems.**



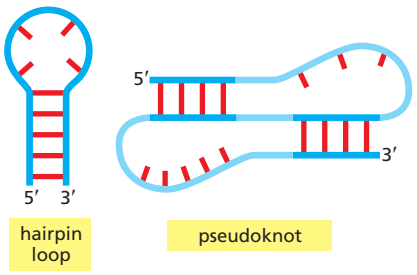


Figure 6–89 **Some common elements of RNA structure.** Conventional, complementary base-pairing interactions are indicated by *red* "rungs" in double-helical portions of the RNA.

The RNA world hypothesis relies on the fact that, among present-day biological molecules, RNA is unique in being able to act as both a carrier of genetic information and as a ribozyme to catalyze chemical reactions. In this section, we discuss these properties of RNA and how they may have been especially important in early cells.

Single-Stranded RNA Molecules Can Fold into Highly Elaborate Structures

We have seen in this chapter that RNA can carry genetic information in mRNAs, and we saw in Chapter 5 that the genomes of some viruses are composed solely of RNA. We have also seen that complementary base-pairing and other types of hydrogen bonds can occur between nucleotides in the same chain of RNA, causing an RNA molecule to fold up in a unique way determined by its nucleotide sequence (see, for example, Figures 6–50 and 6–67). 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 6–89**).

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 a catalyst (**Figure 6–90**). 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 provided by the limited chemical groups of a polynucleotide chain.

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 6–91**). Such experiments have created RNAs that can catalyze a wide variety of biochemical reactions (**Table 6–5**), with reaction rate enhancements only a few orders of magnitude lower than

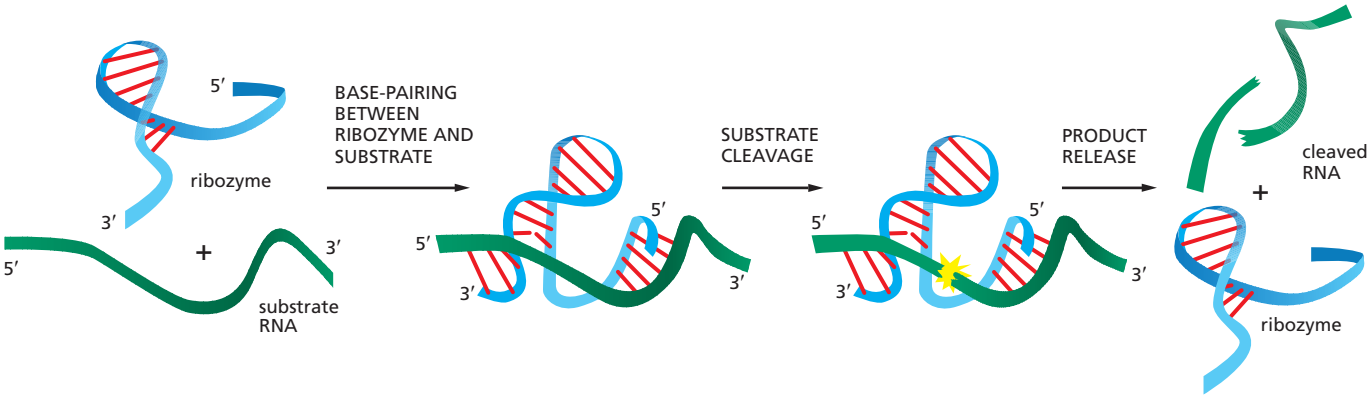


Figure 6–90 **A ribozyme.** This simple RNA molecule catalyzes the cleavage of a second RNA at a specific site. This ribozyme is found embedded in larger RNA genomes—called viroids—which infect plants. The cleavage, which occurs in nature at a distant location on the same RNA molecule that contains the ribozyme, is a step in the replication of the viroid genome. Although not shown in the figure, the reaction requires a magnesium ion 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.)

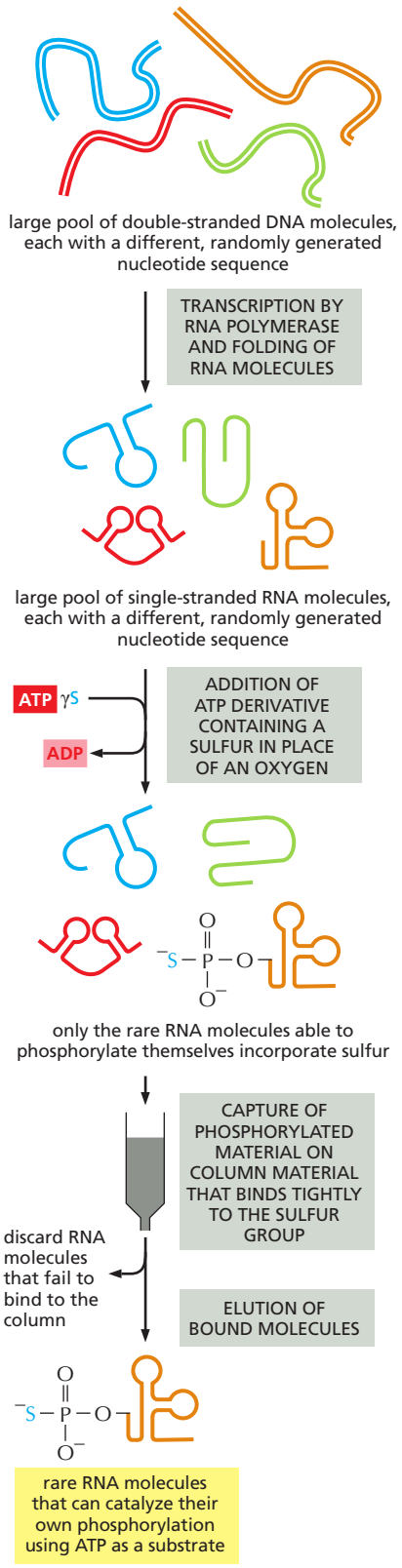


Figure 6–91 ***In vitro*** **selection of a synthetic ribozyme.** Beginning with a large pool of nucleic acid molecules synthesized in the laboratory, those rare RNA molecules that possess a specified catalytic activity can be isolated and studied. Although a specific example (that of an autophosphorylating 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.)

those of the "fastest" protein enzymes. Given these findings, it is not clear why protein catalysts greatly outnumber ribozymes in modern cells. Experiments have shown, however, that RNA molecules may have more difficulty than proteins in binding to flexible, hydrophobic substrates. In any case, the availability of 20 types of amino acids presumably provides proteins with a greater number of catalytic strategies.

RNA Can Both Store Information and Catalyze Chemical Reactions

RNA molecules 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 their nucleotide subunits, which enables one RNA to act as a template for the formation of another. As we have seen in this and the preceding chapter, these complementary templating mechanisms lie at the heart of DNA replication and transcription in modern-day cells.

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, 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

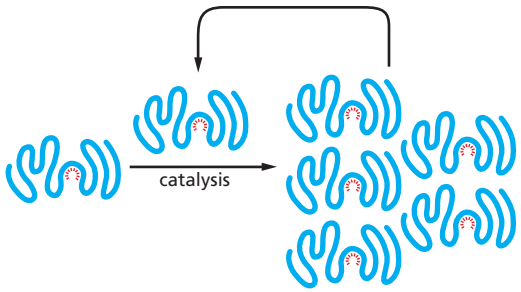


Figure 6–92 An RNA molecule that can catalyze its own synthesis. This hypothetical process would require catalysis both of the production of a second RNA strand of complementary nucleotide sequence (not shown) and the use of this second RNA molecule as a template to form many molecules of RNA with the original sequence. The *red* rays represent the active site of this hypothetical RNA enzyme.

But the efficient synthesis of RNA by such complementary templating mechanisms requires catalysts to promote the polymerization reaction: without catalysts, polymer formation is slow, error-prone, and inefficient.

Because RNA 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 6–92**), it has been proposed that RNAs served long ago as the catalysts for template-dependent RNA synthesis. Although self-replicating systems of RNA molecules have not been found in nature, scientists have made significant progress toward constructing them in the laboratory. While such demonstrations would not prove that self-replicating RNA molecules were central to the origin of life on Earth, they would establish that such a scenario is plausible.

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. As attractive as the RNA world 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 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 noncoded, primitive version of protein synthesis first developed in 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. However, it leaves unexplained how the genetic code—which lies at the core of protein synthesis in today's cells—might have arisen. We 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 adaptors, each matched to a specific amino acid, could have arisen in the RNA world, marking the beginnings of a genetic code.

Once coded protein 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 these ideas are highly speculative, they are consistent with the known properties of RNA and protein molecules.

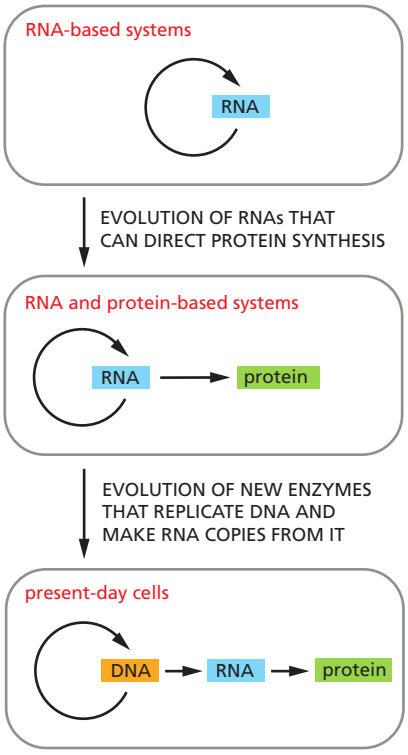


Figure 6–93 The hypothesis that RNA preceded DNA and proteins in evolution. In the earliest cells, RNA molecules (or their close analogs) would have had combined genetic, structural, and catalytic functions. In present-day cells, DNA is the repository of genetic information, and proteins perform the 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.

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 6–93**). 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 pre-dates 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 (pp. 271–273).

Summary

From our knowledge of present-day organisms and the molecules they contain, it seems likely that the development of the distinctive autocatalytic mechanisms fundamental to living systems began with the evolution of families of RNA molecules that could catalyze their own replication. DNA is likely to have been a late addition: as the accumulation of 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.

WHAT WE DON'T KNOW

- How did the present relationships between nucleic acids and proteins evolve? How did the genetic code originate?
- The information carried in genomes specifies the sequences of all proteins and RNA molecules in the cell, and it determines when and where these molecules are synthesized. Do genomes carry other types of information that we have not yet discovered?
- Cells go to great length to correct mistakes in the processes of DNA replication, transcription, splicing, and translation. Are there analogous strategies to correct mistakes in the selection of which genes are to be expressed in a given cell type? Could the great complexity of transcription initiation in animals and plants reflect such a strategy?

PROBLEMS

Which statements are true? Explain why or why not.

6–1 The consequences of errors in transcription are less severe than those of errors in DNA 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 During protein synthesis, the thermodynamics of base-pairing between tRNAs and mRNAs sets the upper limit for the accuracy with which protein molecules are made.

6–5 Protein enzymes are thought to greatly outnumber ribozymes in modern cells because they can catalyze a much greater variety of reactions and all of them have faster rates than any ribozyme.

Discuss the following problems.

6–6 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 6–6).

6–7 You have attached an RNA polymerase molecule to a glass slide and have allowed it to initiate transcription on a template DNA that is tethered to a magnetic bead as shown in **Figure Q6–2**. If the DNA with its attached magnetic bead moves relative to the RNA polymerase as indicated in the figure, in which direction will the bead rotate?

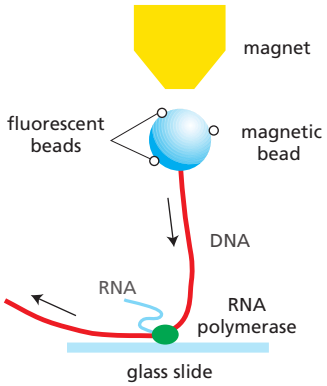


Figure Q6–2 System for measuring the rotation of DNA caused by RNA polymerase (Problem 6–7). The magnet holds the bead upright (but doesn't interfere with its rotation), and the attached tiny fluorescent beads allow the direction of motion to be visualized under the microscope. RNA polymerase is held in place by attachment to the glass slide.

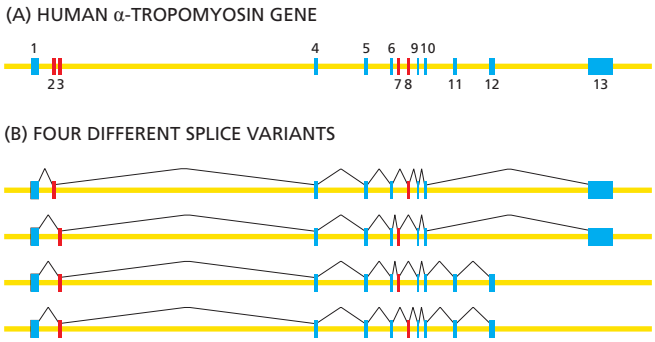


Figure Q6–3 Alternatively spliced mRNAs from the human α-tropomyosin gene (Problem 6–8). (A) Exons in the human α-tropomyosin gene. The locations and relative sizes of exons are shown by the *blue* and *red rectangles*, with alternative exons in *red*. (B) Splicing patterns for four α-tropomyosin mRNAs. Splicing is indicated by *lines* connecting the exons that are included in the mRNA.

6–8 The human α-tropomyosin gene is alternatively spliced to produce different forms of α-tropomyosin mRNA in different cell types (**Figure Q6–3**). For all forms of the mRNA, the protein sequences encoded by exon 1 are the same, as are the protein sequences encoded by exon 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).

6–9 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 Q6–4**). 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 Q6–4). 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?

6–10 Which of the following mutational changes would you predict to be the most deleterious to gene function? Explain your answers.

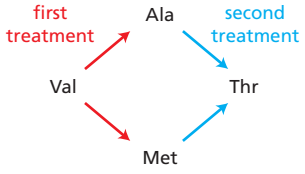


Figure Q6–4 Two rounds of mutagenesis and the altered amino acids at a single position in a protein (Problem 6–9).

1. Insertion of a single nucleotide near the end of the coding sequence.
2. Removal of a single nucleotide near the beginning of the coding sequence.
3. Deletion of three consecutive nucleotides in the middle of the coding sequence.
4. Substitution of one nucleotide for another in the middle of the coding sequence.

6–11 Prokaryotes and eukaryotes both protect against the dangers of translating broken mRNAs. What dangers do partial mRNAs pose for the cell?

6–12 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?

6–13 Most proteins require molecular chaperones to assist in their correct folding. How do you suppose the chaperones themselves manage to fold correctly?

6–14 What is so special about RNA that it is hypothesized to be an evolutionary precursor to DNA and protein? What is it about DNA that makes it a better material than RNA for storage of genetic information?

6–15 If an RNA molecule could form a hairpin with a symmetric internal loop, as shown in **Figure Q6–5**, 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?

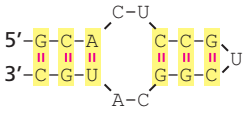


Figure Q6–5 An RNA hairpin with a symmetric internal loop (Problem 6–15).

6–16 Imagine a warm pond on the primordial Earth. Chance processes have just assembled a single copy of an RNA molecule with a catalytic site that can carry out RNA replication. This RNA molecule folds into a structure that is capable of linking nucleotides according to instructions in an RNA template. Given an adequate supply of nucleotides, will this single RNA molecule be able to use itself as a template to catalyze its own replication? Why or why not?

REFERENCES

General

Atkins JF, Gesteland RF & Cech TR (eds) (2011) The RNA Worlds: From Life's Origins to Diversity in Gene Regulation. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Berg JM, Tymoczko JL & Stryer L (2012) Biochemistry, 7th ed. New York: WH Freeman.

Brown TA (2007) Genomes 3. New York: Garland Science.

Darnell J (2011) RNA: Life's Indispensable Molecule. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Hartwell L, Hood L, Goldberg ML et al. (2011) Genetics: from Genes to Genomes, 4th ed. Boston: McGraw Hill.

Judson HF (1996) The Eighth Day of Creation, 25th anniversary ed. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Lodish H, Berk A, Kaiser C et al. (2012) Molecular Cell Biology, 7th 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. (2013) Molecular Biology of the Gene, 7th ed. Menlo Park, CA: Benjamin Cummings.

From DNA to RNA

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.

Brenner S, Jacob F & Meselson M (1961) An unstable intermediate carrying information from genes to ribosomes for protein synthesis. *Nature* 190, 576–581.

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.

Conaway CC & Conaway JW (2011) Function and regulation of the Mediator complex. *Curr. Opin. Genet. Dev.* 21, 225–230.

Cooper TA, Wan L & Dreyfuss G (2009) RNA and disease. *Cell* 136, 777–793.

Cramer P, Armache KJ, Baumli S et al. (2008) Structure of eukaryotic RNA polymerases. *Annu. Rev. Biophys.* 37, 337–352.

Fica SM, Tuttle N, Novak T et al. (2013) RNA catalyses nuclear pre-mRNA splicing. *Nature* 503, 229–234.

Grunberg S & Hahn S (2013) Structural insights into transcription initiation by RNA polymerases II. *Trends Biochem. Sci.* 38, 603–611.

Grunwald D, Singer RH & Rout M (2011) Nuclear export dynamics of TNA-protein complexes. *Nature* 475, 333–341.

Kornblihtt AR, Schor IE, Allo M et al. (2013) Alternative splicing: a pivotal step between eukaryotic transcription and translation. *Nature* 14, 153–165.

Liu X, Bushnell DA & Kornberg RD (2012) RNA polymerase II transcription: Structure and mechanism. *Biochim. Biophys. Acta* 1829, 2–8.

Makino DL, Halibach F & Conti E (2013) The RNA exosome and proteasome: common principles of degradation control. *Nature* 14, 654–660.

Malik S & Roeder RC (2010) The metazoan mediator co-activator complex as an integrative hub for transcriptional regulation. *Nat. Rev. Genet.* 11, 761–772.

Mao YS, Zhang B & Spector DL (2011) Biogenesis and function of nuclear bodies. *Trends Genet.* 27, 295–306.

Matera AG & Wang Z (2014) A day in the life of the spliceosome. *Nature* 15, 108–121.

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.

Opalka N, Brown J, Lane WJ et al. (2010) Complete structural model of *Escherichia coli* RNA polymerase from a hybrid approach. *PLoS Biol.* 9, 1–16.

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.

Schneider C & Tollervey D (2013) Threading the barrel of the RNA exosome. *Trends Biochem. Sci.* 38, 485–493.

Semlow DR & Staley JP (2012) Staying on message: ensuring fidelity in pre-mRNA splicing. *Trends Biochem. Sci.* 37, 263–273.

From RNA to Protein

Anfinsen CB (1973) Principles that govern the folding of protein chains. *Science* 181, 223–230.

Crick FHC (1966) The genetic code: III. *Sci. Am.* 215, 55–62.

Forster F, Unverdorben P, Sledz P et al. (2013) Unveiling the long-held secrets of the 26S proteasome. *Structure* 21, 1551–1562.

Hershko A, Ciechanover A & Varshavsky A (2000) The ubiquitin system. *Nat. Med.* 6, 1073–1081.

Horwich AL, Fenton WA, Chapman E et al. (2007) Two families of chaperonin: physiology and mechanism. *Annu. Rev. Cell Dev. Biol*. 23, 115–145.

Ling J, Reynolds N & Ibba M (2009) Aminoacyl-tRNA synthesis and translational quality control. *Annu Rev. Microbiol.* 63, 61–78.

Moore PB (2012) How should we think about the ribosome? *Annu. Rev. Biophys.* 41, 1–19.

Noller HF (2005) RNA structure: reading the ribosome. *Science* 309, 1508–1514.

Popp MW & Maquat LE (2013) Organizing principles of mammalian nonsense-mediated mRNA decay. *Annu. Rev. Genet.* 47, 139–165.

Saibil H (2013) Chaperone machines for protein folding, unfolding and disaggregation. *Nature* 14, 630–642.

Schmidt M & Finley D (2013) Regulation of proteasome activity in health and disease. *Biochim. Biophys. Acta* 1843, 13–25.

Steitz TA (2008) A structural understanding of the dynamic ribosome machine. *Nature* 9, 242–253.

Varshavsky A (2012) The ubiquitin system, an immense realm. *Annu. Rev. Biochem.* 81, 167–176.

Voorhees RM & Ramakrishnan V (2013) Structural basis of the translational elongation cycle. *Annu. Rev. Biochem.* 82, 203–236.

Wilson DN (2014) Ribosome-targeting antibiotics and mechanisms of bacterial resistance. *Nat. Rev. Microbiol.* 12, 35–48.

Zaher HS & Green R (2009) Fidelity at the molecular level: Lessons from protein synthesis. *Cell* 136, 746–762.

The RNA World and the Origins of Life

Blain JC & Szostak JW (2014) Progress Towards Synthetic Cells. *Annu. Rev. Biochem.* 83, 615–640.

Cech TR (2009) Crawling out of the RNA world. *Cell* 136, 599–602.

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.

Orgel L (2000) Origin of life. A simpler nucleic acid. *Science* 290, 1306–1307.

Robertson MP& Joyce GF (2012) The origins of the RNA world. *Cold Spring Harb. Perspect. Biol.* 4, a003608.

CHAPTER 7

Control of Gene Expression

IN THIS CHAPTER

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 focus on the first half of this problem—the rules and mechanisms that enable a subset of genes to be selectively expressed in each cell. These mechanisms operate at many levels, and we shall discuss each level in turn. But first we present some of the basic principles involved.

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 liver cell, 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 often seemed irreversible, biologists originally suspected that genes might be selectively lost when a cell differentiates. We now know, however, that cell differentiation generally occurs without changes in the nucleotide sequence of a 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. The initial evidence that they do this without altering the sequence of their DNA came 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

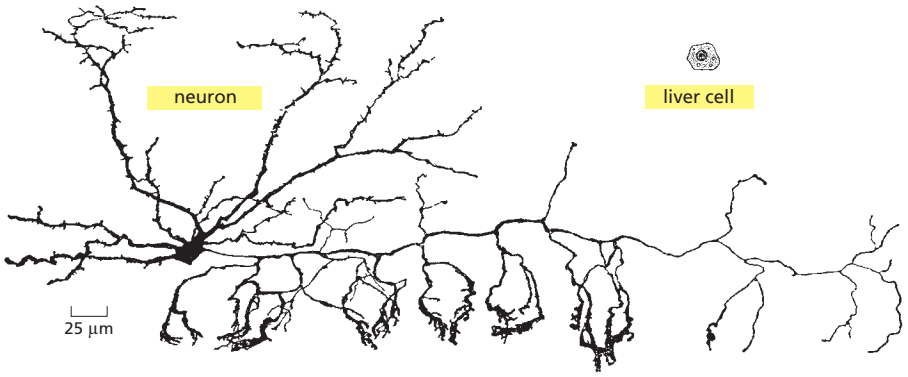


Figure 7–1 A neuron and a liver cell share the same genome. The long branches of this neuron from the retina enable it to receive electrical signals from many other neurons and convey them to neighboring neurons. The liver cell, which is drawn to the same scale, is involved in many metabolic processes, including digestion and the detoxification of alcohol and other drugs. Both of these mammalian cells contain the same genome, but they express different sets of RNAs and proteins. (Neuron adapted from S. Ramón y Cajal, Histologie du Systeme Nerveux de l'Homme et de Vertebres, 1909–1911. Paris: Maloine; reprinted, Madrid: C.S.I.C, 1972.)

produce a normal tadpole (**Figure 7–2**A). The tadpole contains a full range of differentiated cells that derived their DNA sequences from the nucleus of the original donor cell. Thus, the differentiated donor cell cannot have lost any important DNA sequences. A similar conclusion came from experiments performed with plants. When 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). And the same principle has been more recently demonstrated in mammals that include sheep, cattle, pigs, goats, dogs, and mice (Figure 7–2C).

Most recently, detailed DNA sequencing has confirmed the conclusion that the changes in gene expression that underlie the development of multicellular organisms do not generally involve changes in the DNA sequence of the genome.

Different Cell Types Synthesize Different Sets of RNAs and 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

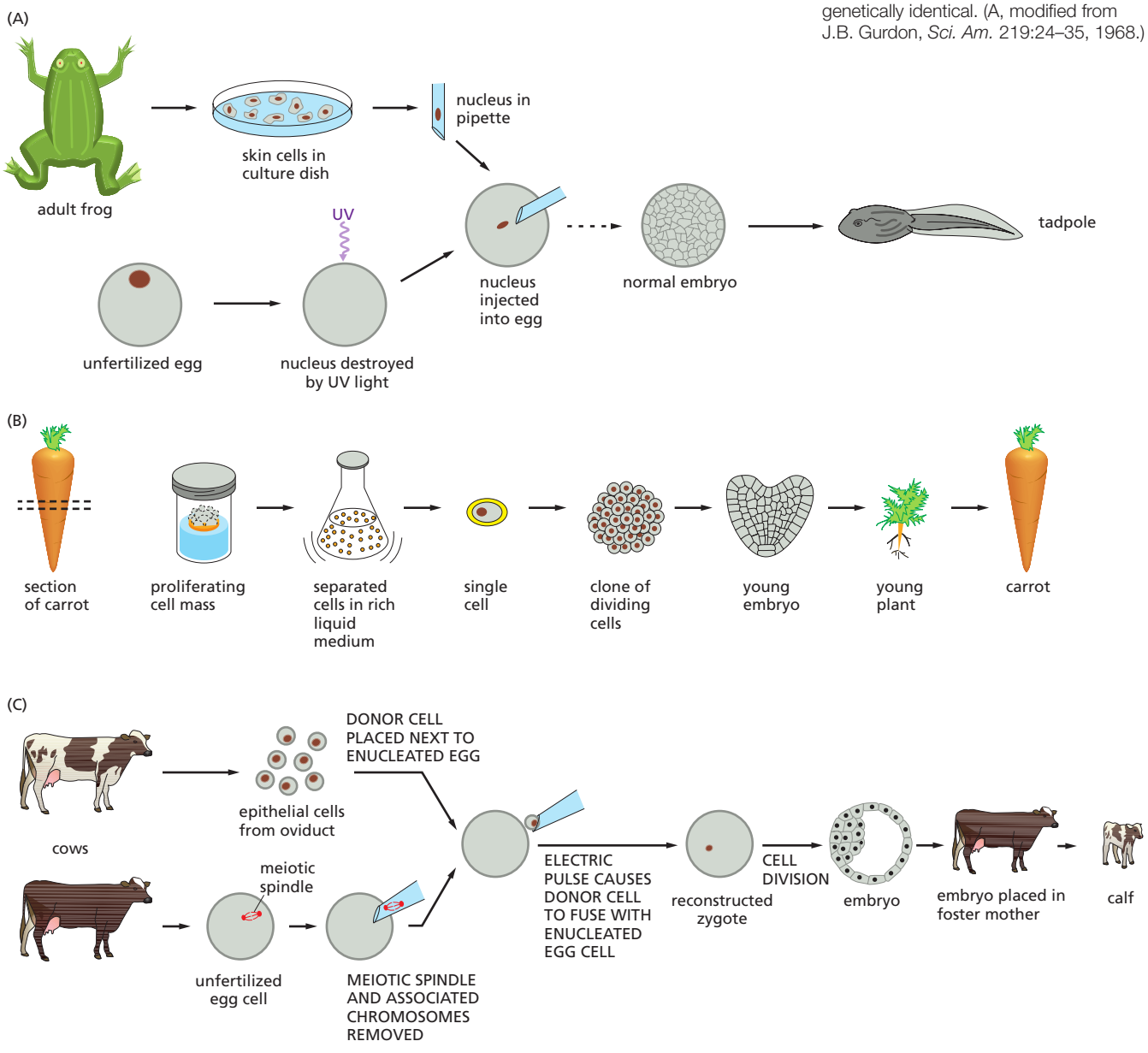


Figure 7–2 Differentiated cells contain all the genetic instructions necessary to direct the formation of a complete organism. (A) The nucleus of a skin cell from an adult frog transplanted into an enucleated egg can give rise to an entire tadpole. The *broken arrow* indicates that, to give the transplanted genome time to adjust to an embryonic environment, a further transfer step is required in which one of the nuclei is taken from an 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 "de-differentiate," so that a single cell can form a clone of progeny cells that later give rise to an entire plant. (C) A nucleus removed from a differentiated cell from an adult cow and 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 all clones of the donor and are therefore genetically identical. (A, modified from J.B. Gurdon, *Sci. Am.* 219:24–35, 1968.)

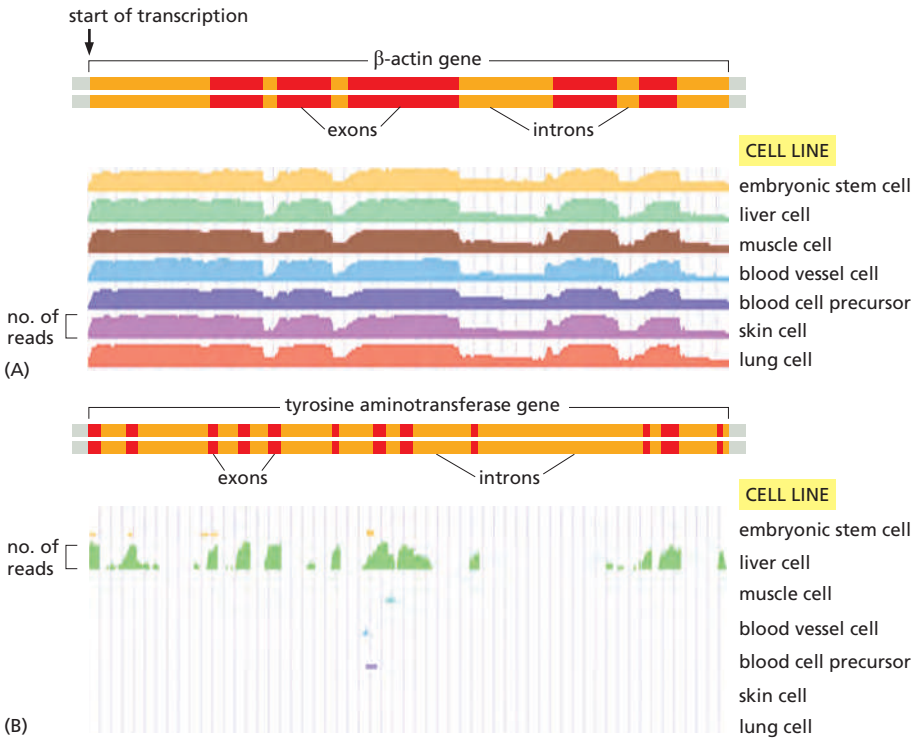


Figure 7–3 Differences in RNA levels for two human genes in seven different tissues. To obtain the RNA data by the technique known as *RNA-seq* (see p. 447), RNA was collected from human cell lines grown in culture, derived from each of the seven indicated tissues. Millions of "sequence reads" were obtained and mapped across the human genome by matching RNA sequences to the DNA sequence of the genome. At each position along the genome, the height of the colored trace is proportional to the number of sequence reads that match the genome sequence at that point. As seen in the figure, the exon sequences in transcribed genes are present at high levels, reflecting their presence in mature mRNAs. Intron sequences are present at much lower levels and reflect pre-mRNA molecules that have not yet been spliced plus intron sequences that have been spliced out but not yet degraded. (A) The gene coding for "all-purpose" actin, a major component of the cytoskeleton. Note that the left-hand end of the mature β-actin mRNA is not translated into protein. As explained later in this chapter, many mRNAs have 5′ untranslated regions that regulate their translation into protein. (B) The same type of data displayed for the enzyme tyrosine aminotransferase, which is highly expressed in liver cells but not in the other cell types tested. (Information for both panels from the University of California, Santa Cruz, Genome Browser (http://genome.ucsc.edu), which provides this type of information for every human gene. See also S. Djebali et al., *Nature* 489:101–108, 2012.)

still do not have an exact answer to this fundamental question, we can make several general statements.

1. Many processes are common to all cells, and any two cells in a single organism therefore have many gene products in common. These include the structural proteins of chromosomes, RNA and DNA polymerases, DNA repair enzymes, ribosomal proteins and RNAs, the enzymes that catalyze the central reactions of metabolism, and many of the proteins that form the cytoskeleton such as actin (**Figure 7–3**A).
2. Some RNAs and proteins are abundant in the specialized cells in which they function and cannot be detected elsewhere, even by sensitive tests. Hemoglobin, for example, is expressed specifically in red blood cells, where it carries oxygen, and the enzyme tyrosine aminotransferase (which breaks down tyrosine in food) is expressed in liver but not in most other tissues (Figure 7–3B).
3. Studies of the number of different RNAs suggest that, at any one time, a typical human cell expresses 30–60% of its approximately 30,000 genes at some level. There are about 21,000 protein-coding genes and a roughly estimated 9000 noncoding RNA genes in humans. When the patterns of RNA expression in different human cell lines are compared, the level of expression of almost every gene is found to vary from one cell type to another. A few of these differences are striking, like those of hemoglobin and tyrosine aminotransferase noted above, but most are much more subtle. But even those genes that are expressed in all cell types usually vary in their *level* of expression from one cell type to the next.
4. Although there are striking differences in coding RNAs (mRNAs) in specialized cell types, they underestimate the full range of differences in the final pattern of protein production. As we discuss in this chapter, there are many steps after RNA production at which gene expression can be regulated. And, as we saw in Chapter 3, proteins are often covalently modified after they are synthesized. The radical differences in gene expression between cell types are therefore most fully revealed through methods that directly display the levels of proteins along with their post-translational modifications (**Figure 7–4**).

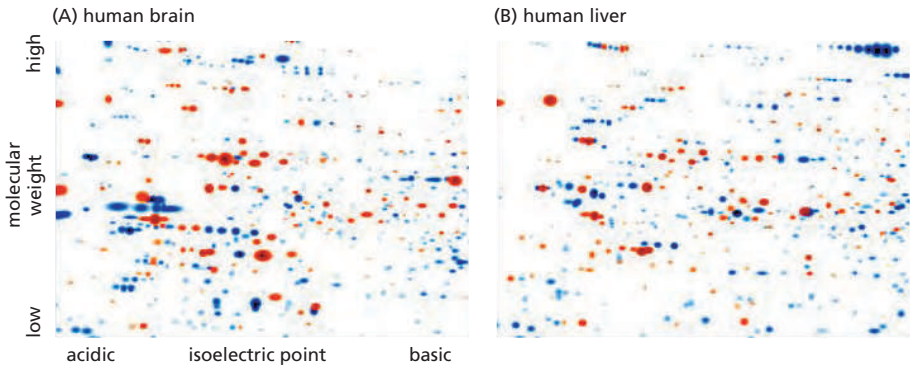


Figure 7–4 Differences in the proteins expressed by two human tissues, (A) brain and (B) liver. In each panel, the proteins are displayed using two-dimensional polyacrylamide-gel electrophoresis (see pp. 452–454). The proteins have been separated by molecular weight (*top* to *bottom*) and isoelectric 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 that tissue. 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 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.

Methods based on mass spectrometry (see pp. 455–457) provide much more detailed information, including the identity of each protein, the position of each modification, and the nature of the modification. (Courtesy of Tim Myers and Leigh Anderson, Large Scale Biology Corporation.)

External Signals Can Cause a Cell to Change the Expression of Its Genes

Although the specialized cells in a multicellular organism have characteristic patterns of gene expression, each cell is capable of altering its pattern of gene expression in response to extracellular cues. If a liver cell is exposed to a glucocorticoid hormone, for example, the production of a set of proteins is dramatically increased. Released in the body during periods of starvation or intense exercise, glucocorticoids signal the liver to increase the production of energy from amino acids and other small molecules; the set of proteins whose production is induced includes the enzyme tyrosine aminotransferase, mentioned above. When the hormone is no longer present, the production of these proteins drops to its normal, unstimulated level in liver cells.

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 very differently to the same extracellular signal. Other features of the gene expression pattern 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 localizing 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 then return to the additional ways of regulating gene expression.

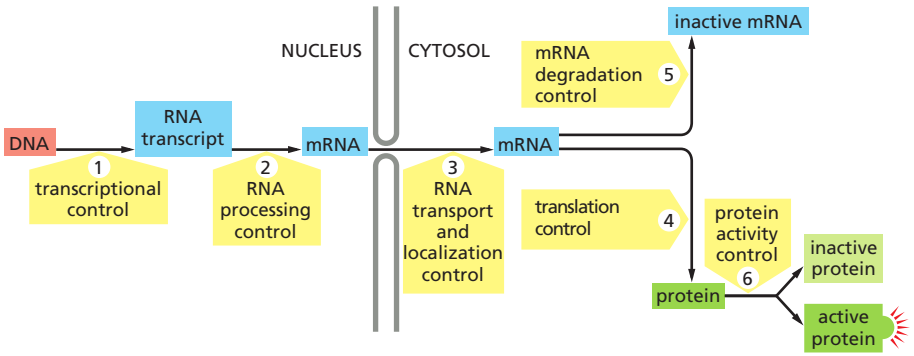


Figure 7–5 Six steps at which eukaryotic gene expression can be controlled. Controls that operate at steps 1 through 5 are discussed in this chapter. Step 6, the regulation of protein activity, occurs largely through covalent post-translational modifications including phosphorylation, acetylation, and ubiquitylation (see Table 3–3, p. 165). Step 6 was introduced in Chapter 3 and is subsequently discussed in many chapters throughout the book.

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 provides the most important point of control.

CONTROL OF TRANSCRIPTION BY SEQUENCE-SPECIFIC DNA-BINDING PROTEINS

How does a cell determine which of its thousands of genes to transcribe? Perhaps the most important concept, one that applies to all species on Earth, is based on a group of proteins known as **transcription regulators**. These proteins recognize specific sequences of DNA (typically 5–10 nucleotide pairs in length) that are often called ***cis*-regulatory sequences**, because they must be on the same chromosome (that is, *in cis*) to the genes they control. Transcription regulators bind to these sequences, which are dispersed throughout genomes, and this binding puts into motion a series of reactions that ultimately specify which genes are to be transcribed and at what rate. Approximately 10% of the protein-coding genes of most organisms are devoted to transcription regulators, making them one of the largest classes of proteins in the cell. In most cases, a given transcription regulator recognizes its own *cis*-regulatory sequence, which is different from those recognized by all the other regulators in the cell.

Transcription of each gene is, in turn, controlled by its own collection of *cis*-regulatory sequences. These typically lie near the gene, often in the intergenic region directly upstream from the transcription start point of the gene. Although a few genes are controlled by a single *cis*-regulatory sequence that is recognized by a single transcription regulator, the majority have complex arrangements of *cis*-regulatory sequences, each of which is recognized by a different transcription regulator. It is therefore the positions, identity, and arrangement of *cis*-regulatory sequences—which are an important part of the information embedded in the genome—that ultimately determine the time and place that each gene is transcribed.

We begin our discussion by describing how transcription regulators recognize *cis*-regulatory sequences.

The Sequence of Nucleotides in the DNA Double Helix Can Be Read by Proteins

As discussed in Chapter 4, the DNA in a chromosome consists of a very long double helix that has both a major and a minor groove (**Figure 7–6**). Transcription regulators must recognize short, specific *cis*-regulatory sequences within

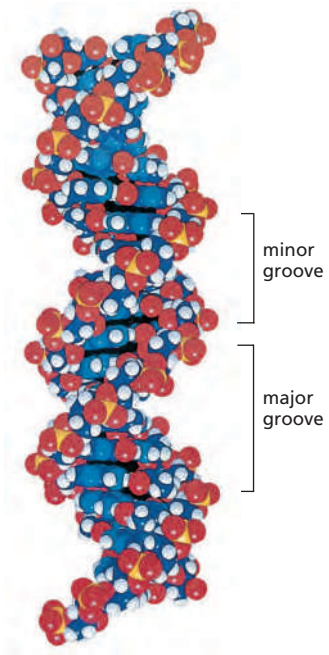
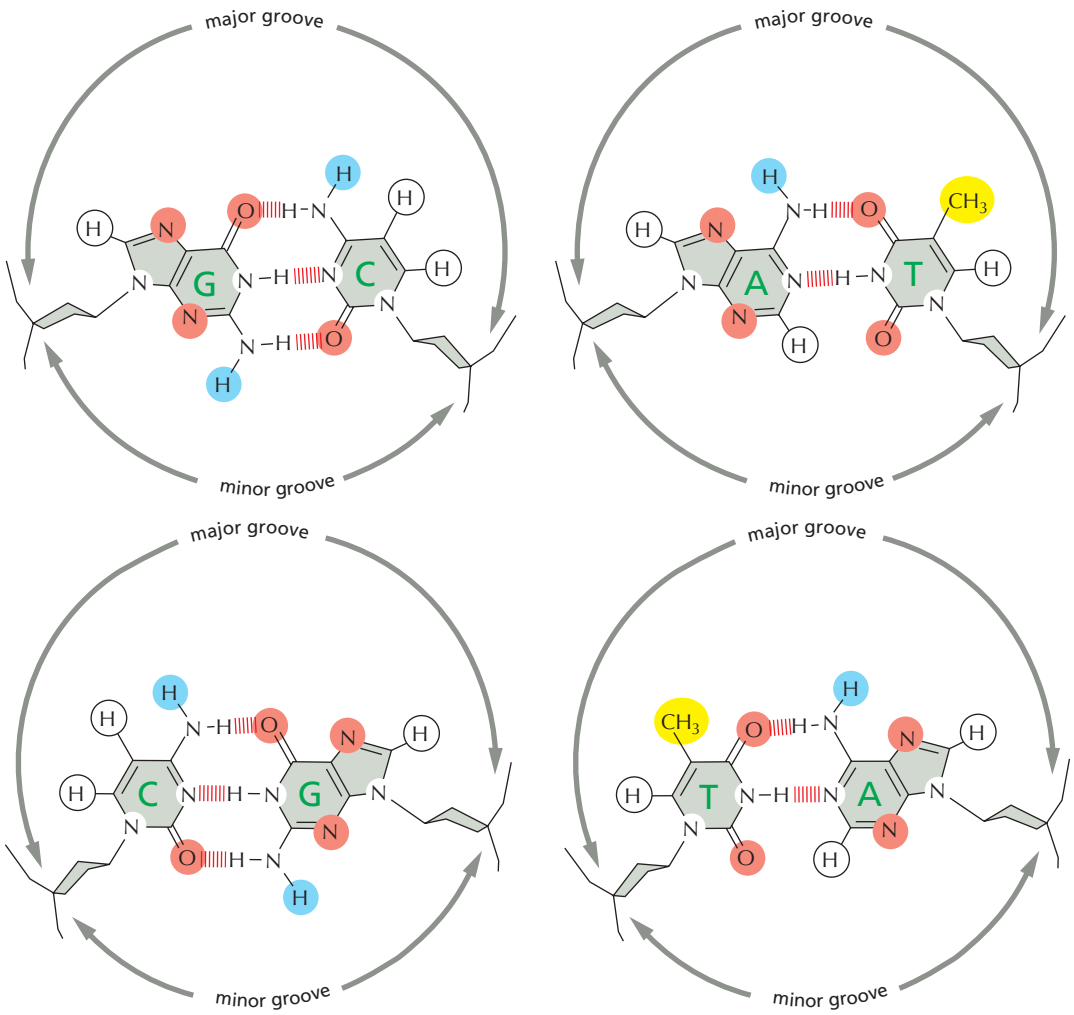


Figure 7–6 Double-helical structure of DNA. A space-filling model of DNA showing the major and minor grooves on the outside of the double helix (see Movie 4.1). The atoms are colored as follows: carbon, *dark blue*; nitrogen, *light blue*; hydrogen, *white*; oxygen, *red*; phosphorus, *yellow*.



this structure. When first discovered in the 1960s, it was thought that these proteins might require direct access to the interior of the double helix to distinguish between one DNA sequence and another. It is now clear, however, that the outside of the double helix is studded with DNA sequence information that transcription regulators recognize: the edge of each base pair presents a distinctive pattern of hydrogen-bond donors, hydrogen-bond acceptors, and hydrophobic patches in both the major and minor grooves (**Figure 7–7**). Because the major groove is wider and displays more molecular features than does the minor groove, nearly all transcription regulators make the majority of their contacts with the major groove—as we shall see.

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 the major groove, each of the four base-pair configurations projects a unique pattern of features. (From C. Branden and J. Tooze, Introduction to Protein Structure, 2nd ed. New York: Garland Publishing, 1999.)

Transcription Regulators 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 transcription regulators has provided some of the clearest examples of this principle. A transcription regulator recognizes a specific *cis*-regulatory sequence because the surface of the protein is extensively complementary to the special surface features of the double helix that displays that sequence. Each transcription regulator 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 contacts that are typically formed at the protein–DNA interface add together to ensure that the interaction is both highly specific and very strong (**Figure 7–8**). In fact, DNA–protein

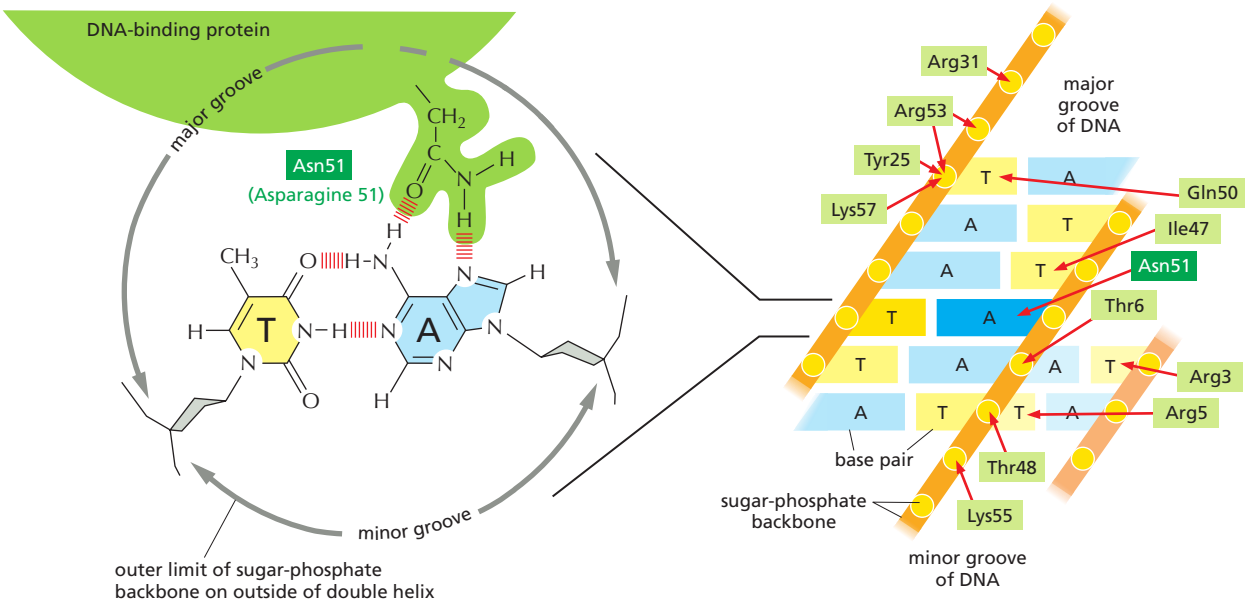


Figure 7–8 The binding of a transcription regulator to a specific DNA sequence. On the *left*, a single contact is shown between a transcription regulator and DNA; such contacts allow the protein to "read" the DNA sequence. On the *right*, the complete set of contacts between a transcription regulator (a member of the homeodomain family—see Panel 7–1) and its *cis*-regulatory sequence is shown. The DNA-binding portion of the protein is 60 amino acids long. Although the interactions in the major groove are the most important, the protein is also seen to contact both the minor groove and phosphates in the sugar–phosphate DNA backbone. (See C. Wolberger et al., *Cell* 67:517–528, 1991.)

interactions include some of the tightest and most specific molecular interactions known in biology.

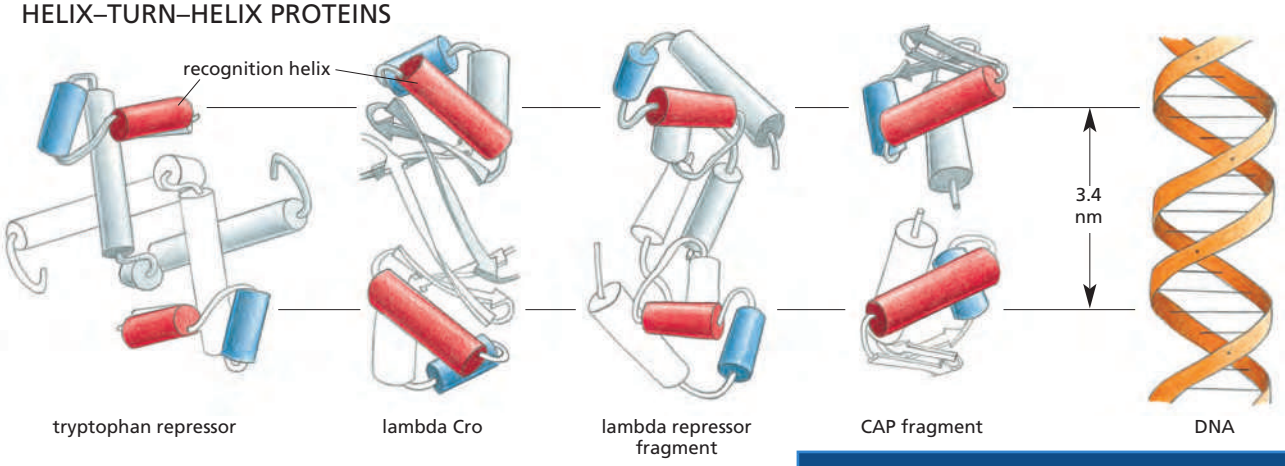
Although each example of protein–DNA recognition is unique in detail, x-ray crystallographic and nuclear magnetic resonance (NMR) spectroscopic studies of hundreds of transcription regulators have revealed that many of them contain one or another of a small set of DNA-binding structural motifs (**Panel 7–1**). These motifs generally use either α helices or β sheets to bind to the major groove of DNA. The amino acid side chains that extend from these protein motifs make the specific contacts with the DNA. Thus, a given structural motif can be used to recognize many different *cis*-regulatory sequences depending on the specific side chains present.

Dimerization of Transcription Regulators Increases Their Affinity and Specificity for DNA

A monomer of a typical transcription regulator recognizes about 6–8 nucleotide pairs of DNA. However, sequence-specific DNA-binding proteins do not bind tightly to a single DNA sequence and reject all others; rather, they recognize a range of closely related sequences, with the affinity of the protein for the DNA varying according to how closely the DNA matches the optimal sequence. Hence, *cis*-regulatory sequences are often depicted as "logos" which display the range of sequences recognized by a particular transcription regulator (**Figure 7–9**A and B). In Chapter 6, we saw this same representation at work for the binding of RNA polymerase to promoters (see Figure 6–12).

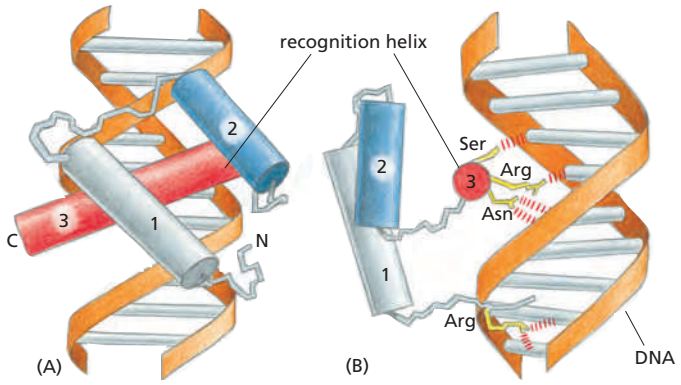
The DNA sequence recognized by a monomer does not contain sufficient information to be picked out from the background of such sequences that would occur at random all over the genome. For example, an exact six-nucleotide DNA sequence would be expected to occur by chance approximately once every 4096 nucleotides (4^6), and the range of six-nucleotide sequences described by a typical logo would be expected to occur by chance much more often, perhaps every 1000 nucleotides. Clearly, for a bacterial genome of 4.6×10^6 nucleotide pairs, not to mention a mammalian genome of 3×10^9 nucleotide pairs, this is insufficient information to accurately control the transcription of individual genes. Additional contributions to DNA-binding specificity must therefore be present. Many transcription regulators form dimers, with both monomers making nearly identical contacts with DNA (Figure 7–9C). This arrangement doubles the length of the *cis*-regulatory sequence recognized and greatly increases both the affinity and the specificity of transcription regulator binding. Because the DNA sequence

HELIX–TURN–HELIX PROTEINS



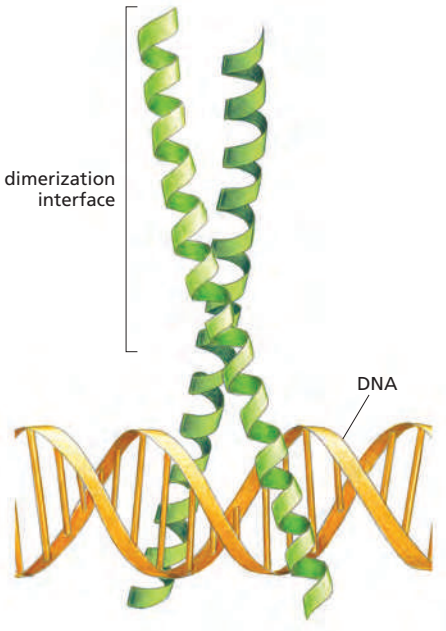
Originally identified in bacterial transcription regulators, this motif has since been found in many hundreds of DNA-binding proteins from both eukaryotes and prokaryotes. It is constructed from two α helices (*blue* and *red*) connected by a short extended chain of amino acids, which constitutes the "turn." The two helices are held at a fixed angle, primarily through interactions between the two helices. The more C-terminal helix (in *red*) 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. All of the proteins shown here bind DNA as dimers in which the two copies of the recognition helix (in *red*) are separated by exactly one turn of the DNA helix (3.4 nm); thus both recognition helices of the dimer can fit into the major groove of DNA.

HOMEODOMAIN PROTEINS



Not long after the first transcription regulators were discovered in bacteria, genetic analyses of 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 (discussed in Chapter 21). It was later shown that these genes coded for transcription regulators that bound DNA through a structural motif named the homeodomain. 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 7–8. A flexible arm attached to helix 1 forms contacts with nucleotide pairs in the minor groove.

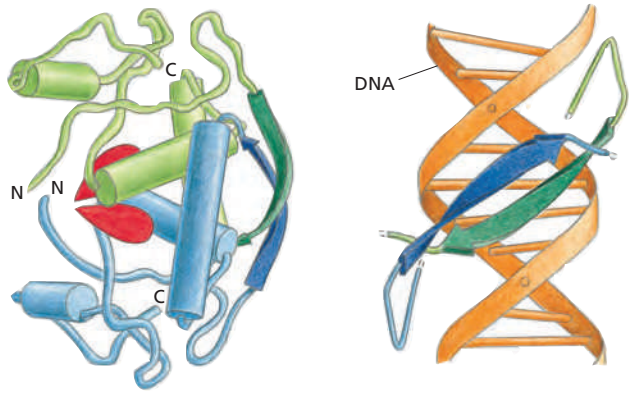
LEUCINE ZIPPER PROTEINS



The *leucine zipper* motif is named because of the way the two α helices, one from each monomer, are joined together to form a short coiled-coil. These proteins bind DNA as dimers where the two long α 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.

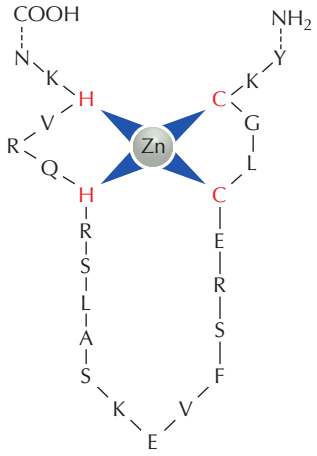
β SHEET DNA RECOGNITION PROTEINS

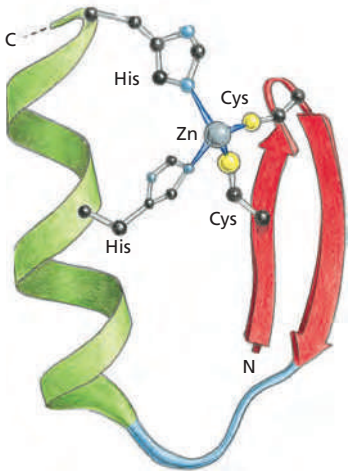
In the other DNA-binding motifs displayed in this panel, α helices are the primary mechanism used to recognize specific DNA sequences. In one large group of transcription regulators, however, a two-stranded β sheet, with amino acid side chains extending from the sheet toward the DNA, reads the information on the surface of the major groove. 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. Shown is a transcription regulator that binds two molecules of *S*-adenosyl methionine (*red*). On the left is a dimer of the protein; on the right is a simplified diagram showing just the two-stranded β sheet bound to the major groove of DNA.

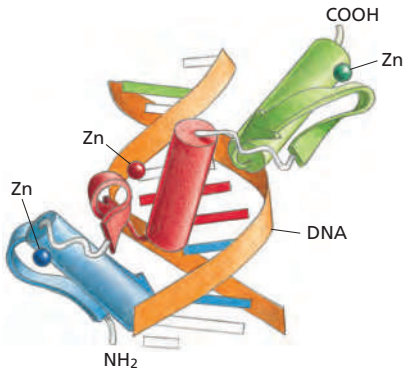


ZINC FINGER PROTEINS

This group of DNA-binding motifs includes one or more zinc atoms as structural components. All such zinc-coordinated DNA-binding motifs are called zinc fingers, referring to their appearance in early schematic drawings (*left*). They fall into several distinct structural groups, only one of which we consider here. It has a simple structure, in which the zinc atom holds an α helix and a β sheet together (*middle*). This type of zinc finger is often found in clusters with the α helix of each finger contacting the major groove of the DNA, forming a nearly continuous stretch of α helices along that groove. In this way, a strong and specific DNA–protein interaction is built up through a repeating basic structural unit. Three such fingers are shown on the *right*.

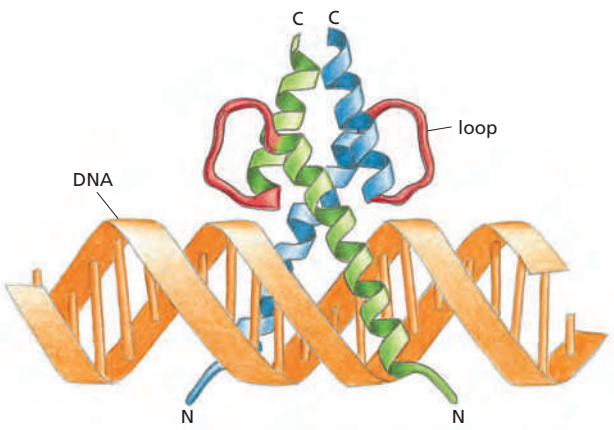






HELIX–LOOP–HELIX PROTEINS

Related to the leucine zipper, the helix–loop–helix motif consists of a short α helix connected by a loop (*red*) to a second, longer α helix. The flexibility of the loop allows one helix to fold back and park against the other thereby forming the dimerization surface. As shown, this two-helix structure binds both to DNA and to the two-helix structure of a second protein to create either a homodimer or a heterodimer. Two α helices that extend from the dimerization interface make specific contacts with the major groove of DNA.



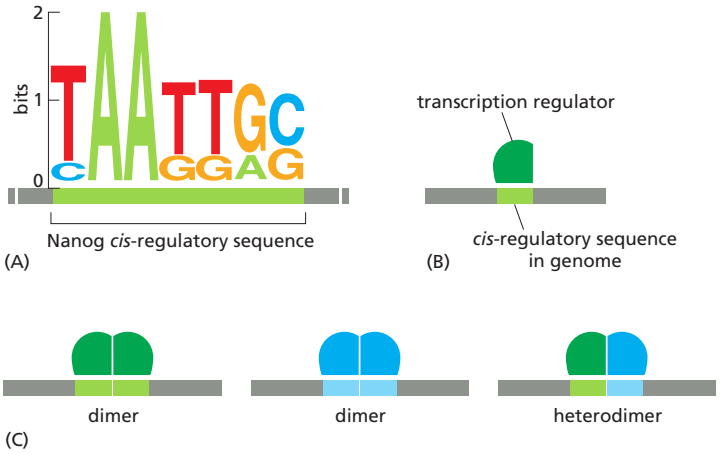


Figure 7–9 Transcription regulators and *cis*-regulatory sequences. (A) Depiction of the *cis*-regulatory sequence for Nanog, a homeodomain family member that is a key regulator in embryonic stem cells. This "logo" form (see Figure 6–12) shows that the protein can recognize a collection of closely related DNA sequences and gives the preferred nucleotide pair at each position. *Cis*-regulatory sequences are "read" as double-stranded DNA, but only one strand typically is shown in a logo. (B) Representation of the *cis*-regulatory sequence as a colored box. (C) Many transcription regulators form dimers (homodimers) and heterodimers. In the example shown, three different DNA-binding specificities are formed from two transcription regulators.

recognized by the protein has increased from approximately 6 nucleotide pairs to 12 nucleotide pairs, there are many fewer random occurrences of matching sequences.

Heterodimers are often formed from two different transcription regulators. Transcription regulators may form heterodimers with more than one partner protein; in this way, the same transcription regulator can be "reused" to create several distinct DNA-binding specificities (see Figure 7–9C).

Transcription Regulators Bind Cooperatively to DNA

In the simplest case, the collection of noncovalent bonds that holds the above dimers or heterodimers together is so extensive that these structures form obligatorily, and never fall apart. In this case, the unit of binding is the dimer or heterodimer, and the binding curve for the transcription regulator (the fraction of DNA bound as a function of protein concentration) has a standard exponential shape (**Figure 7–10**A).

In many cases, however, the dimers and heterodimers are held together very weakly; they exist predominantly as monomers in solution, and yet dimers are observed on the appropriate DNA sequence. Here, the proteins are said to bind to DNA cooperatively, and the curve describing their binding is sigmoidal in shape (Figure 7–10B). *Cooperative binding* means that, over a range of concentrations of the transcription regulator, binding is more of an all-or-none phenomenon than

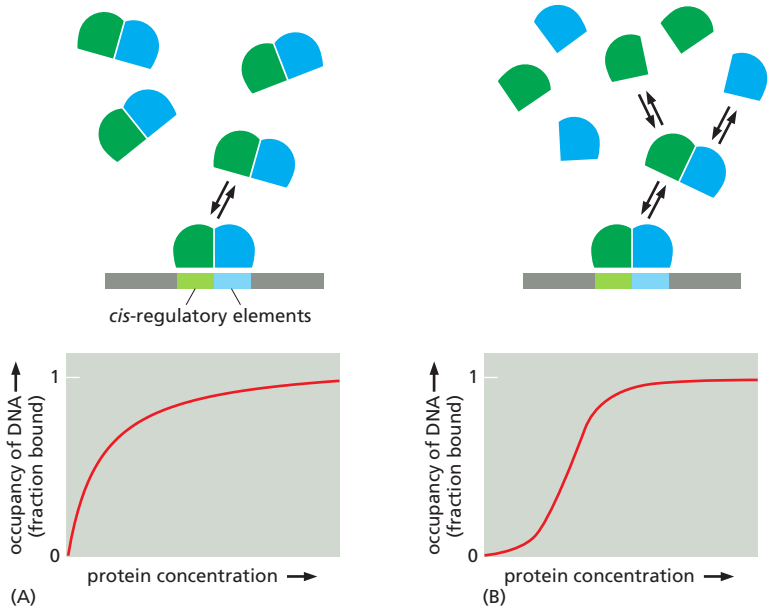


Figure 7–10 Occupancy of a *cis*-regulatory sequence by a transcription regulator. (A) Noncooperative binding by a stable heterodimer. (B) Cooperative binding by components of a heterodimer that are predominantly monomers in solution. The shape of the curve differs from that of (A) because the fraction of protein in a form competent to bind DNA (the heterodimer) increases with increasing protein concentration.

for noncooperative binding; that is, at most protein concentrations, the *cis*-regulatory sequence is either nearly empty or nearly fully occupied and rarely is somewhere in between. A discussion of the mathematics behind cooperative binding is given in Chapter 8 (see Figure 8–79A).

Nucleosome Structure Promotes Cooperative Binding of Transcription Regulators

As we have just seen, cooperative binding of transcription regulators to DNA often occurs because the monomers have only a weak affinity for each other. However, there is a second, indirect mechanism for cooperative binding, one that arises from the nucleosome structure of eukaryotic chromosomes.

In general, transcription regulators bind to DNA in nucleosomes with lower affinity than they do to naked DNA. There are two reasons for this difference. First, the surface of the *cis*-regulatory sequence recognized by the transcription regulator may be facing inward on the nucleosome, toward the histone core, and therefore not be readily available to the regulatory protein. Second, even if the face of the *cis*-regulatory sequence is exposed on the outside of the nucleosome, many transcription regulators subtly alter the conformation of the DNA when they bind, and these changes are generally opposed by the tight wrapping of the DNA around the histone core. For example, many transcription regulators induce a bend or kink in the DNA when they bind.

We saw in Chapter 4 that nucleosome remodeling can alter the structure of the nucleosome, allowing transcription regulators access to the DNA. Even without remodeling, however, transcription regulators can still gain limited access to DNA in a nucleosome. The DNA at the end of a nucleosome "breathes," transiently exposing the DNA and allowing regulators to bind. This breathing happens at a much lower rate in the middle of the nucleosome; therefore, the positions where the DNA exits the nucleosome are much easier to occupy (**Figure 7–11**).

These properties of the nucleosome promote cooperative DNA binding by transcription regulators. If a regulatory protein enters the DNA of a nucleosome and prevents the DNA from tightly rewrapping around the nucleosome core, it will increase the affinity of a second transcription regulator for a nearby *cis*-regulatory sequence. If the two transcription regulators also interact with each other (as described above), the cooperative effect is even greater. In some cases, the combined action of the regulatory proteins can eventually displace the histone core of the nucleosome altogether.

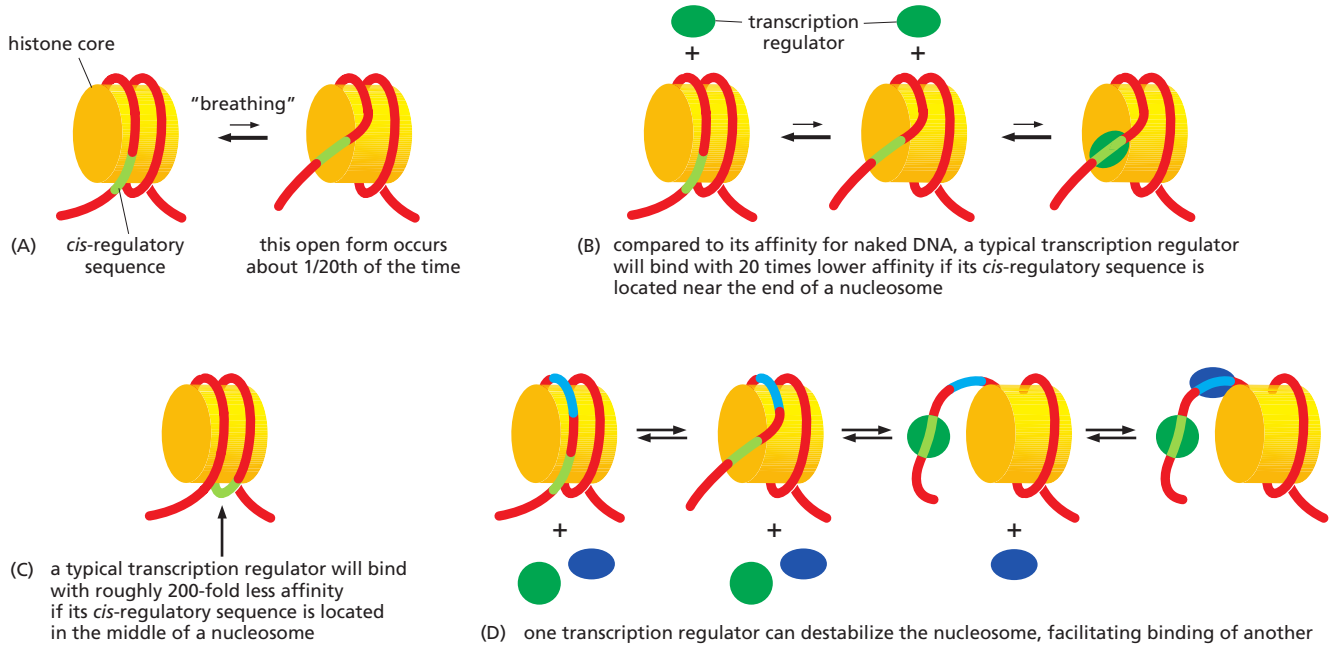


Figure 7–11 How nucleosomes effect the binding of transcription regulators.

The cooperation among transcription regulators can become much greater when nucleosome remodeling complexes are involved. If one transcription regulator binds its *cis*-regulatory sequence and attracts a chromatin remodeling complex, the localized action of the remodeling complex can allow a second transcription regulator to efficiently bind nearby. Moreover, we have discussed how transcription regulators can work together in pairs; in reality, larger numbers often cooperate by repeated use of the same principles. A highly cooperative binding of transcription regulators to DNA probably explains why many sites in eukaryotic genomes that are bound by transcription regulators are "nucleosome free."

Summary

Transcription regulators recognize short stretches of double-helical DNA of defined sequence called cis*-regulatory sequences, and thereby determine which of the thousands of genes in a cell will be transcribed. Approximately 10% of the protein-coding genes in most organisms produce transcription regulators, and they control many features of cells. Although each of these transcription regulators has unique features, most bind to DNA as homodimers or heterodimers and recognize DNA through one of a small number of structural motifs. Transcription regulators typically work in groups and bind DNA cooperatively, a feature that has several underlying mechanisms, some of which exploit the packaging of DNA in nucleosomes.*

TRANSCRIPTION REGULATORS SWITCH GENES ON AND OFF

Having seen how transcription regulators bind to *cis*-regulatory sequences embedded in the genome, we can now discuss how, once bound, these proteins influence the transcription of genes. The situation in bacteria is simpler than in eukaryotes (for one thing, chromatin structure is not an issue), and we therefore discuss it first. Following this, we turn to the more complex situation in eukaryotes.

The Tryptophan Repressor Switches Genes Off

The genome of the bacterium *E. coli* consists of a single, circular DNA molecule of about 4.6×10^6 nucleotide pairs. This DNA encodes approximately 4300 proteins, although only a fraction of these are made at any one time. Bacteria regulate the expression of many of their genes according to the food sources that are available in the environment. For example, in *E. coli*, five genes code for enzymes that manufacture the amino acid tryptophan. These genes are arranged in a cluster on the chromosome and are transcribed from a single promoter as one long mRNA molecule; such coordinately transcribed clusters are called *operons* (**Figure 7–12**). Although operons are common in bacteria, they are rare in eukaryotes, where genes are typically transcribed and regulated individually (see Figure 7–3).

When tryptophan concentrations are low, the operon is transcribed; the resulting mRNA is translated to produce a full set of biosynthetic enzymes, which work in tandem to synthesize tryptophan from much simpler molecules. When tryptophan is abundant, however—for example, when the bacterium is in the gut of a mammal that has just eaten a protein-rich meal—the amino acid is imported into the cell and shuts down production of the enzymes, which are no longer needed.

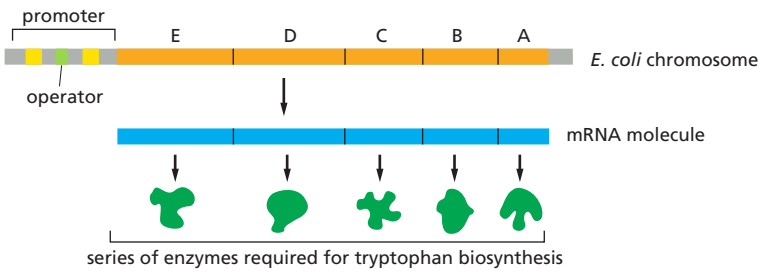


Figure 7–12 A cluster of bacterial genes can be transcribed from a single promoter. Each of these five genes encodes a different enzyme, and all of these enzymes are needed to synthesize the amino acid tryptophan from simpler molecules. The genes are transcribed as a single mRNA molecule, a feature that allows their expression to be coordinated. Clusters of genes transcribed as a single mRNA molecule are common in bacteria. Each of these clusters is called an *operon* because its expression is controlled by a *cis*-regulatory sequence called the *operator (green)*, situated within the promoter. (In this and subsequent figures, the *yellow* blocks in the promoter represent DNA sequences that bind RNA polymerase; see Figure 6–12).

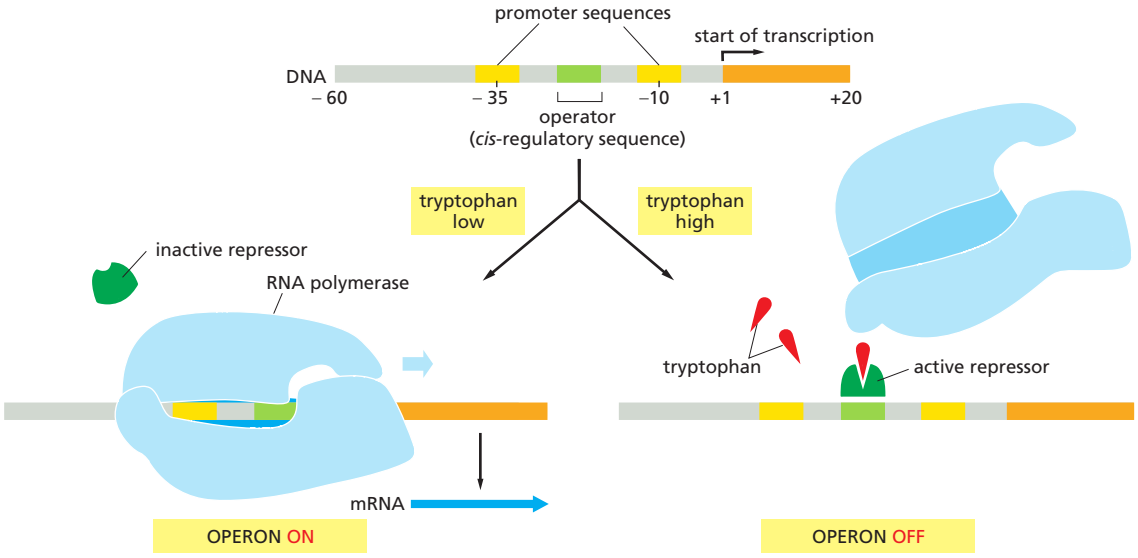


Figure 7–13 Genes can be switched off by repressor proteins. If the concentration of tryptophan inside a bacterium is low *(left),* RNA polymerase *(blue)* binds to the promoter and transcribes the five genes of the tryptophan operon. However, if the concentration of tryptophan is high *(right)*, the repressor protein *(dark green)* becomes active and binds to the operator *(light green),* where it blocks the binding of RNA polymerase to the promoter. Whenever the concentration of intracellular tryptophan drops, the repressor falls off the DNA, allowing the polymerase to again transcribe the operon. Although not shown in the figure, the repressor is a stable dimer.

We now understand exactly how this repression of the tryptophan operon comes about. Within the operon's promoter is a *cis*-regulatory sequence that is recognized by a transcription regulator. When this regulator binds to this sequence, it blocks access of RNA polymerase to the promoter, thereby preventing transcription of the operon (and thus production of the tryptophan-producing enzymes). The transcription regulator is known as the *tryptophan repressor* and its *cis*-regulatory sequence is called the *tryptophan operator*. These components are controlled in a simple way: the repressor can bind to DNA only if it has also bound several molecules of tryptophan (**Figure 7–13**).

The tryptophan repressor is an allosteric protein, and the binding of tryptophan causes a subtle change in its three-dimensional structure so that the protein can bind to the operator sequence. Whenever the concentration of free tryptophan in the bacterium drops, tryptophan dissociates from the repressor, the repressor no longer binds to DNA, and the tryptophan operon is transcribed. The repressor is thus a simple device that switches production of a set of biosynthetic enzymes on and off according to the availability of the end product of the pathway that the enzymes catalyze.

The tryptophan repressor protein itself is always present in the cell. The gene that encodes it is continuously transcribed at a low level, so that a small amount of the repressor protein is always being made. Thus the bacterium can respond very rapidly to a rise or fall in tryptophan concentration.

Repressors Turn Genes Off and Activators Turn Them On

The tryptophan repressor, as its name suggests, is a *transcriptional repressor* protein: in its active form, it switches genes off, or *represses* them. Some bacterial transcription regulators do the opposite: they switch genes on, or *activate* them. These *transcriptional activator* proteins work on promoters that—in contrast to the promoter for the tryptophan operon—are only marginally able to bind and position RNA polymerase on their own. However, these poorly functioning promoters can be made fully functional by activator proteins that bind to nearby *cis*-regulatory sequences and contact the RNA polymerase to help it initiate transcription (**Figure 7–14**).

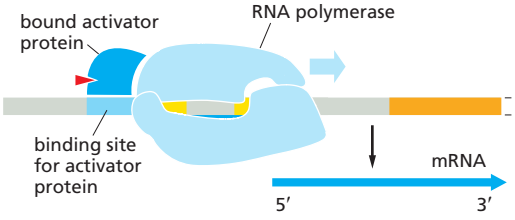


Figure 7–14 Genes can be switched on by activator proteins. An activator protein binds to its *cis*-regulatory sequence on the DNA and interacts with the RNA polymerase to help it initiate transcription. Without the activator, the promoter fails to initiate transcription efficiently. In bacteria, the binding of the activator to DNA is often controlled by the interaction of a metabolite or other small molecule *(red triangle)* with the activator protein. The *Lac* operon works in this manner, as we discuss shortly.

DNA-bound activator proteins can increase the rate of transcription initiation as much as 1000-fold, a value consistent with a relatively weak and nonspecific interaction between the transcription regulator 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 ≈18 kJ/mole, which could be accounted for by just a few weak, noncovalent bonds. Thus, many activator proteins work simply by providing a few favorable interactions that help to attract RNA polymerase to the promoter. To provide this assistance, however, the activator protein must be bound to its *cis*-regulatory sequence, and this sequence must be positioned, with respect to the promoter, so that the favorable interactions can occur.

Like the tryptophan repressor, activator proteins often have to interact with a second molecule to be able to bind DNA. For example, the bacterial activator protein *CAP* has to bind cyclic AMP (cAMP) before it can bind to DNA. Genes activated by CAP are switched on in response to an increase in intracellular cAMP concentration, which rises when glucose, the bacterium's preferred carbon source, is no longer available; as a result, CAP drives the production of enzymes that allow the bacterium to digest other sugars.

An Activator and a Repressor Control the *Lac* Operon

In many instances, the activity of a single promoter is controlled by several different transcription regulators. The *Lac* operon in *E. coli,* for example, is controlled by both the *Lac* repressor and the CAP activator that we just discussed. The *Lac* operon encodes proteins required to import and digest the disaccharide lactose. In the absence of glucose, the bacterium makes cAMP, which activates CAP to switch on genes that allow the cell to utilize alternative sources of carbon—including lactose. It would be wasteful, however, for CAP to induce expression of the *Lac* operon if lactose itself were not present. Thus the Lac repressor shuts off the operon in the absence of lactose. This arrangement enables the control region of the *Lac* operon to integrate two different signals, so that the operon is highly expressed only when two conditions are met: glucose must be absent and lactose must be present (**Figure 7–15**). This genetic circuit thus behaves much like

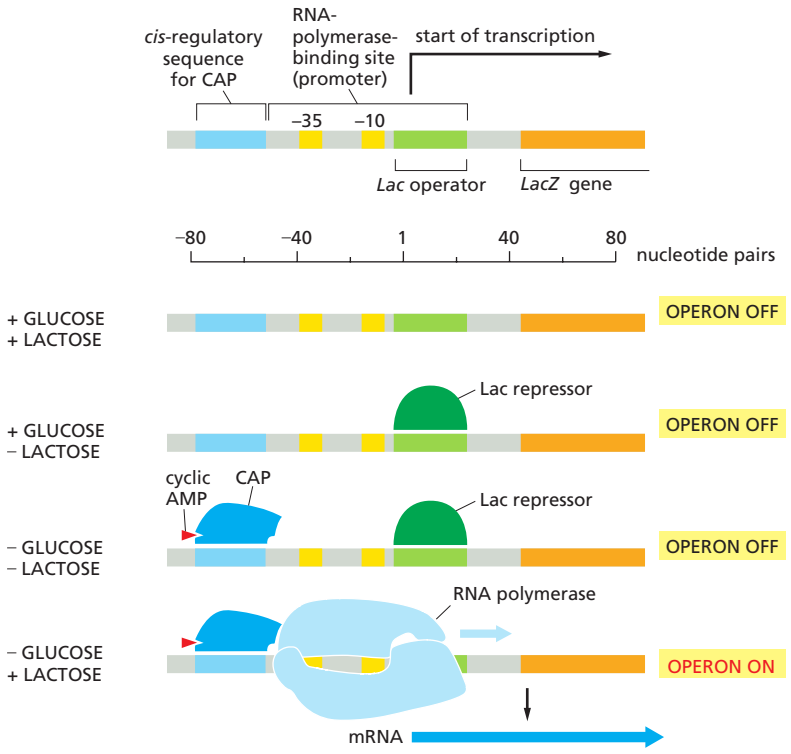


Figure 7–15 The *Lac* operon is controlled by two transcription regulators, the Lac repressor and CAP. *LacZ*, the first gene of the operon, encodes the enzyme β-galactosidase, which breaks down lactose to galactose and glucose. When lactose is absent, the Lac repressor binds to a *cis*-regulatory sequence, called the *Lac* operator, and shuts off expression of the operon (**Movie 7.4**). Addition of lactose increases the intracellular concentration of a related compound, allolactose; allolactose binds to the Lac repressor, causing it to undergo a conformational change that releases its grip on the operator DNA (not shown). When glucose is absent, cyclic AMP *(red triangle)* is produced by the cell, and CAP binds to DNA.

a switch that carries out a logic operation in a computer. When lactose is present AND glucose is absent, the cell executes the appropriate program—in this case, transcription of the genes that permit the uptake and utilization of lactose.

All transcription regulators, whether they are repressors or activators, must be bound to DNA to exert their effects. In this way, each regulatory protein acts selectively, controlling only those genes that bear a *cis*-regulatory sequence recognized by it. The logic of the *Lac* operon first attracted the attention of biologists more than 50 years ago. The way it works was uncovered by a combination of genetics and biochemistry, providing some of the first insights into how transcription is controlled in any organism.

DNA Looping Can Occur During Bacterial Gene Regulation

We have seen that transcription activators help RNA polymerase initiate transcription and repressors hinder it. However, the two types of proteins are very similar to one another. For example, to occupy their *cis*-regulatory sequences, both the tryptophan repressor and the CAP activator protein must bind a small molecule; moreover, they both recognize their *cis*-regulatory sequences using the same structural motif (the helix–turn–helix shown in Panel 7–1). Indeed, some proteins (for example, the CAP protein) can act as both a repressor and an activator, depending on the exact placement of their *cis*-regulatory sequence relative to the promoter: for some genes, the CAP *cis*-regulatory sequence overlaps the promoter, and CAP binding thereby prevents the assembly of RNA polymerase at the promoter.

Most bacteria have small, compact genomes, and the *cis*-regulatory sequences that control the transcription of a gene are typically located very near to the start point of transcription. But there are some exceptions to this generalization—*cis*-regulatory sequences can be located hundreds and even thousands of nucleotide pairs from the bacterial genes they control (**Figure 7–16**). In these cases, the intervening DNA is looped out, allowing a protein bound at a distant site along the DNA to contact RNA polymerase. Here, the DNA acts as a tether, enormously increasing the probability that the proteins will collide, compared with the situation where one protein is bound to DNA and the other is free in solution. We will see shortly that, although it is the exception in bacteria, DNA looping occurs in the regulation of nearly every eukaryotic gene.

A possible explanation for this difference is based on evolutionary considerations. It has been proposed that the compact, simple genetic switches found in bacteria evolved in response to large population sizes where competition for growth put selective pressure on bacteria to maintain small genome sizes. In contrast, there appears to have been little selective pressure to "streamline" the genomes of multicellular organisms.

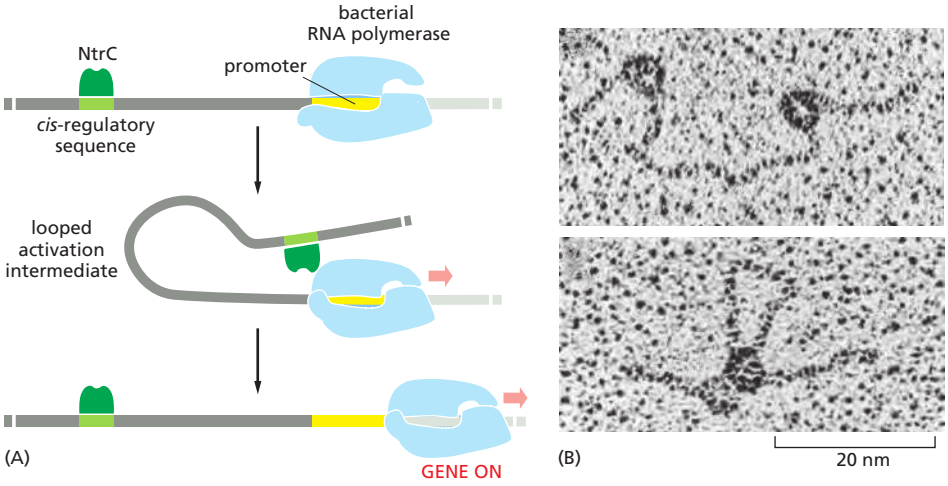


Figure 7–16 Transcriptional activation at a distance. (A) The NtrC protein is a bacterial transcription regulator that activates transcription by directly contacting RNA polymerase. (B) The interaction of NtrC and RNA polymerase, with the intervening DNA looped out, can be seen in the electron microscope. (B, courtesy of Harrison Echols and Sydney Kustu.)

Complex Switches Control Gene Transcription in Eukaryotes

When compared to the situation in bacteria, transcription regulation in eukaryotes involves many more proteins and much longer stretches of DNA. It often seems bewilderingly complex. Yet many of the same principles apply. As in bacteria, the time and place that each gene is to be transcribed is specified by its *cis*-regulatory sequences, which are "read" by the transcription regulators that bind to them. Once bound to DNA, positive transcription regulators (activators) help RNA polymerase begin transcribing genes, and negative regulators (repressors) block this from happening. In bacteria, as we have seen, most of the interactions between DNA-bound transcription regulators and RNA polymerases (whether they activate or repress transcription) are direct. In contrast, these interactions are almost always indirect in eukaryotes: many intermediate proteins, including the histones, act between the DNA-bound transcription regulator and RNA polymerase. Moreover, in multicellular organisms, it is common for dozens of transcription regulators to control a single gene, with *cis*-regulatory sequences spread over tens of thousands of nucleotide pairs. DNA looping allows the DNA-bound regulatory proteins to interact with each other and ultimately with RNA polymerase at the promoter. Finally, because nearly all of the DNA in eukaryotic organisms is compacted by nucleosomes and higher-order structures, transcription initiation in eukaryotes must overcome this inherent block.

In the next sections, we discuss these features of transcription initiation in eukaryotes, emphasizing how they provide extra levels of control not found in bacteria.

A Eukaryotic Gene Control Region Consists of a Promoter Plus Many *cis*-Regulatory Sequences

In eukaryotes, RNA polymerase II transcribes all the protein-coding genes and many noncoding RNA genes, as we saw in Chapter 6. This polymerase requires five general transcription factors (27 subunits *in toto*; see Table 6–3, p. 311), in contrast to bacterial RNA polymerase, which needs only a single general transcription factor (the σ subunit). As we have seen, the stepwise assembly of the general transcription factors at a eukaryotic promoter provides, in principle, multiple steps at which the cell can speed up or slow down the rate of transcription initiation in response to transcription regulators.

Because the many *cis*-regulatory sequences that control the expression of a typical 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 eukaryotic gene. This includes the **promoter**, where the general transcription factors and the polymerase assemble, plus all of the ***cis*-regulatory sequences** to which transcription regulators bind to control the rate of the assembly processes at the promoter (**Figure 7–17**). In animals and plants, it is not unusual to find the regulatory sequences of a gene dotted over stretches of DNA as large as 100,000 nucleotide pairs. Some of this DNA is transcribed (but not translated), and we discuss these long noncoding RNAs (lncRNAs) later in this chapter. For now, we can regard much of this DNA as "spacer" sequences that transcription regulators do not directly recognize. It is important to keep in mind that, like other regions of eukaryotic chromosomes, most of the DNA in gene control regions is packaged into nucleosomes and higher-order forms of chromatin, thereby compacting its overall length and altering its properties.

In this chapter, we shall loosely use the term **gene** to refer to a segment of DNA that is transcribed into a functional RNA molecule, one that either codes for a protein or has a different role in the cell (see Table 6–1, p. 305). However, the classical view of a gene includes the gene control region as well, since mutations in it can produce an altered phenotype. Alternative RNA splicing further complicates the definition of a gene—a point we shall return to later.

In contrast to the small number of *general transcription factors*, which are abundant proteins that assemble on the promoters of all genes transcribed by

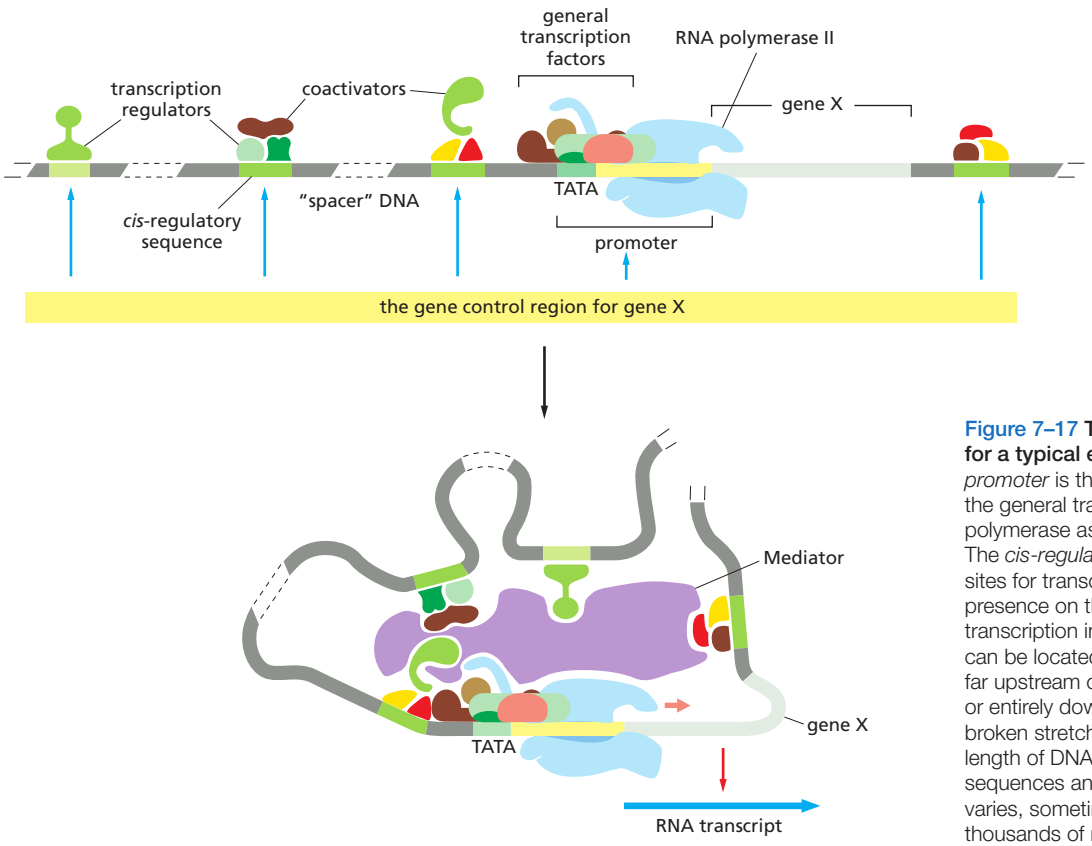


Figure 7–17 The gene control region for a typical eukaryotic gene. The *promoter* is the DNA sequence where the general transcription factors and the polymerase assemble (see Figure 6–15). The *cis-regulatory sequences* are binding sites for transcription regulators, whose presence on the DNA affects the rate of transcription initiation. These sequences can be located adjacent to the promoter, far upstream of it, or even within introns or entirely downstream of the gene. The broken stretches of DNA signify that the length of DNA between the *cis*-regulatory sequences and the start of transcription varies, sometimes reaching tens of thousands of nucleotide pairs in length. The *TATA box* is a DNA recognition sequence for the general transcription factor TFIID. As shown in the lower panel, DNA looping allows transcription regulators bound at any of these positions to interact with the proteins that assemble at the promoter. Many transcription regulators act through Mediator (described in Chapter 6), while some interact with the general transcription factors and RNA polymerase directly. Transcription regulators also act by recruiting proteins that alter the chromatin structure of the promoter (not shown, but discussed below).

Whereas Mediator and the general transcription factors are the same for all RNA polymerase II-transcribed genes, the transcription regulators and the locations of their binding sites relative to the promoter differ for each gene.

RNA polymerase II, there are thousands of different *transcription regulators* devoted to turning individual genes on and off. In eukaryotes, operons—sets of genes transcribed as a unit—are rare, and, instead, each gene is regulated individually. Not surprisingly, the regulation of each gene is different in detail from that of every other gene, and it is difficult to formulate simple rules for gene regulation that apply in every case. We can, however, make some generalizations about how transcription regulators, once bound to gene control regions on DNA, set in motion the series of events that lead to gene activation or repression.

Eukaryotic Transcription Regulators Work in Groups

In bacteria, we saw that proteins such as the tryptophan repressor, the *Lac* repressor, and the CAP protein bind to DNA on their own and directly affect RNA polymerase at the promoter. Eukaryotic transcription regulators, in contrast, usually assemble in groups at their *cis*-regulatory sequences. Often two or more regulators bind cooperatively, as discussed earlier in the chapter. In addition, a broad class of multisubunit proteins termed *coactivators* and *co-repressors* assemble on DNA with them. Typically, these coactivators and co-repressors do not recognize specific DNA sequences themselves; they are brought to those sequences by the transcription regulators. Often the protein–protein interactions between transcription regulators and between regulators and coactivators are too weak for them to assemble in solution; however, the appropriate combination of *cis*-regulatory sequences can "crystallize" the assembly of these complexes on DNA (**Figure 7–18**).

As their names imply, coactivators are typically involved in activating transcription and co-repressors in repressing it. In the following sections, we will see that coactivators and co-repressors can act in a variety of different ways to influence transcription after they have been localized on the genome by transcription regulators.

As shown in Figure 7–18, an individual transcription regulator can often participate in more than one type of regulatory complex. A protein might function,

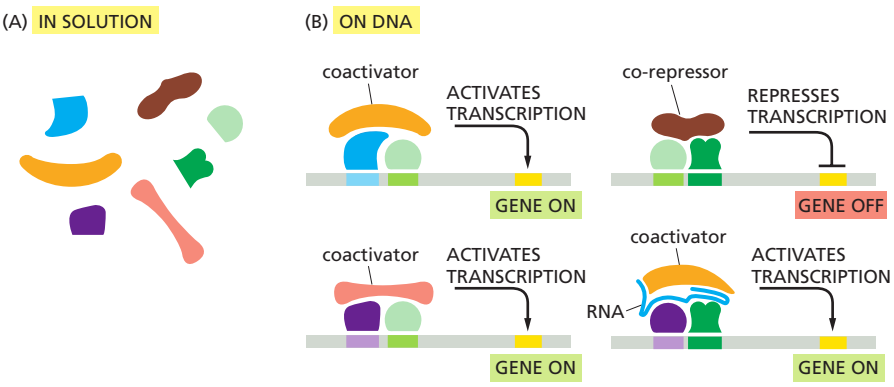


Figure 7–18 Eukaryotic transcription regulators assemble into complexes on DNA. (A) Seven transcription regulators are shown. The nature and function of the complex they form depend on the specific *cis*-regulatory sequences that seed their assembly. (B) Some assembled complexes activate gene transcription, while another represses transcription. Note that the *light green* and *dark green* proteins are shared by both activating and repressing complexes. Proteins that do not themselves bind DNA but assemble on other DNA-bound transcription regulators are termed coactivators or co-repressors. In some cases *(lower right)*, RNA molecules are found in these assemblies. As described later in this chapter, these RNAs often act as scaffolds to hold a group of proteins together.

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 eukaryotic transcription regulators function as regulatory parts that are used to build complexes whose function depends on the final assembly of all of the individual components. Each eukaryotic gene is therefore regulated by a "committee" of proteins, all of which must be present to express the gene at its proper level.

Activator Proteins Promote the Assembly of RNA Polymerase at the Start Point of Transcription

The *cis*-regulatory sequences to which eukaryotic transcription activator proteins bind were originally called *enhancers* because their presence "enhanced" the rate of transcription initiation. It came as a surprise when it was discovered that these sequences could be found tens of thousands of nucleotide pairs away from the promoter; as we have seen, DNA looping, which was not widely appreciated at the time, can now explain this initially puzzling observation.

Once bound to DNA, how do assemblies of activator proteins increase the rate of transcription initiation? At most genes, mechanisms work in concert. Their function is both to attract and position RNA polymerase II at the promoter and to release it so that transcription can begin.

Some activator proteins bind directly to one or more of the general transcription factors, accelerating their assembly on a promoter that has been brought in proximity—through DNA looping—to that activator. Most transcription activators, however, attract coactivators that then perform the biochemical tasks needed to initiate transcription. One of the most prevalent coactivators is the large *Mediator* protein complex, composed of more than 30 subunits. About the same size as RNA polymerase itself, Mediator serves as a bridge between DNA-bound transcription activators, RNA polymerase, and the general transcription factors, facilitating their assembly at the promoter (see Figure 7–17).

Eukaryotic Transcription Activators Direct the Modification of Local Chromatin Structure

The eukaryotic general transcription factors and RNA polymerase are unable, on their own, to assemble on a promoter that is packaged in nucleosomes. Thus, in addition to directing the assembly of the transcription machinery at the promoter, eukaryotic transcription activators promote transcription by triggering changes to the chromatin structure of the promoters, making the underlying DNA more accessible.

The most important ways of locally altering chromatin are through covalent histone modifications, nucleosome remodeling, nucleosome removal, and histone replacement (discussed in Chapter 4). Eukaryotic transcription activators use all four of these mechanisms: thus they attract coactivators that include histone modification enzymes, ATP-dependent chromatin remodeling complexes, and histone chaperones, each of which can alter the chromatin structure of

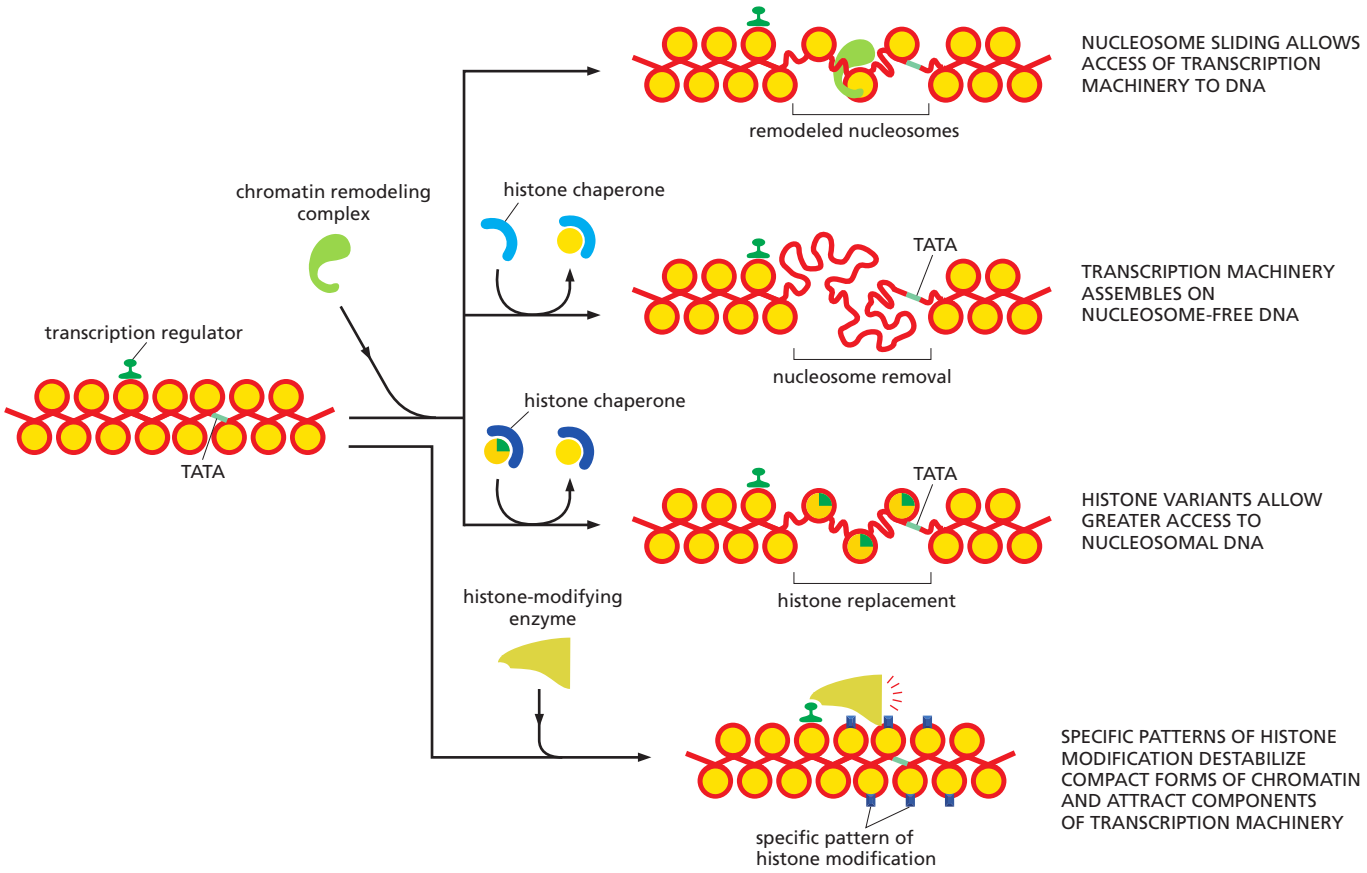


Figure 7–19 Eukaryotic transcription activator proteins direct local alterations in chromatin structure. Nucleosome remodeling, nucleosome removal, histone replacement, and certain types of histone modifications favor transcription initiation (see Figure 4–39). These alterations increase the accessibility of DNA and facilitate the binding of RNA polymerase and the general transcription factors.

promoters (**Figure 7–19**). These local alterations in chromatin structure provide greater access to DNA, thereby facilitating the assembly of the general transcription factors at the promoter. In addition, some histone modifications specifically attract these proteins to the promoter. These mechanisms often work together during transcription initiation (**Figure 7–20**). Finally, as discussed earlier in this chapter, the local chromatin changes directed by one transcriptional regulator can allow the binding of additional regulators. By repeated use of this principle, large assemblies of proteins can form on control regions of genes to regulate their transcription.

The alterations of chromatin structure that occur during transcription initiation can persist for different lengths of time. In some cases, as soon as the transcription regulator 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. In other cases, the altered chromatin structure persists, even after the transcription regulator 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 4–44). The fact that different histone modifications persist for different times provides the cell with a mechanism that makes possible both longer- and shorter-term memory of gene expression patterns.

A special type of chromatin modification occurs as RNA polymerase II transcribes through a gene. The histones just ahead of the polymerase can be acetylated by enzymes carried by the polymerase, removed by histone chaperones, and deposited behind the moving polymerase. These histones are then rapidly deacetylated and methylated, also by complexes that are carried by the polymerase, leaving behind nucleosomes that are especially resistant to transcription. This remarkable process seems to prevent spurious transcription reinitiation

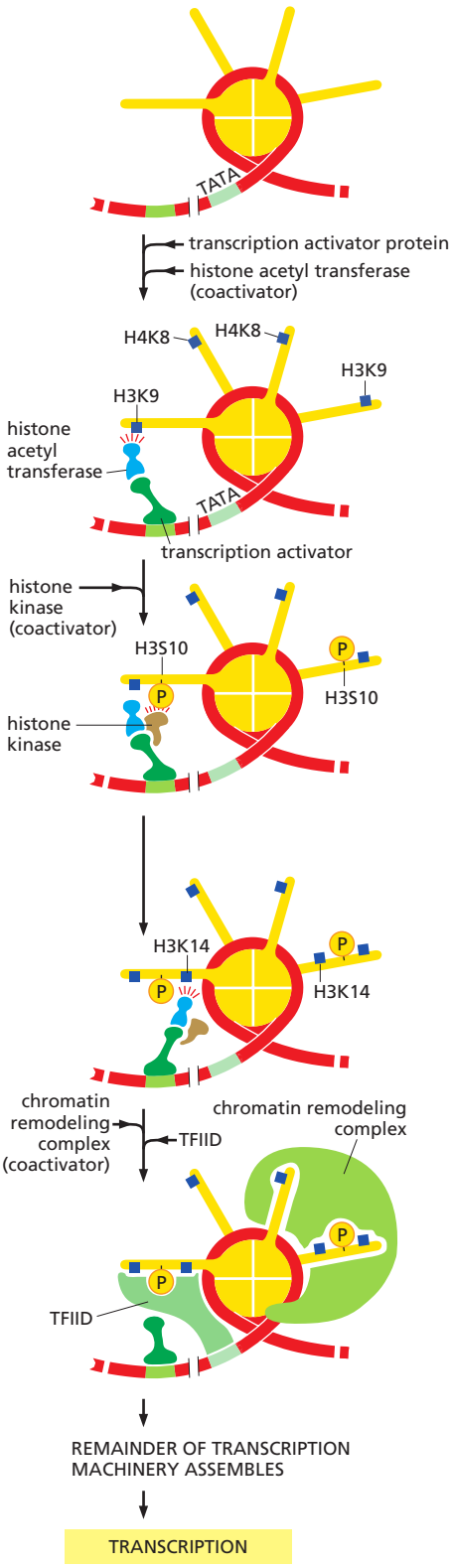


Figure 7–20 Successive histone modifications during transcription initiation. In this example, taken from the human interferon gene promoter, a transcription activator binds to DNA packaged into chromatin and attracts a histone acetyl transferase that acetylates lysine 9 of histone H3 and lysine 8 of histone H4. Then a histone kinase, also attracted by the transcription activator, phosphorylates serine 10 of histone H3 but it can only do so after lysine 9 has been acetylated. This serine modification signals the histone acetyl transferase to acetylate position K14 of histone H3. Next, the general transcription factor TFIID and a chromatin remodeling complex bind to the chromatin to promote the subsequent steps of transcription initiation. TFIID and the remodeling complex 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.

The histone acetyl transferase, the histone kinase, and the chromatin remodeling complex are all coactivators. The order of events shown applies to a specific promoter; at other genes, the steps may occur in a different order or individual steps may be omitted altogether. (Adapted from T. Agalioti, G. Chen and D. Thanos, *Cell* 111:381–392, 2002. With permission from Elsevier.)

behind a moving polymerase, which, in essence, must clear a path through chromatin as it transcribes. Later in this chapter, when we discuss *RNA interference*, the potential dangers to the cell of such inappropriate transcription will become especially obvious. The modification of nucleosomes behind a moving RNA polymerase also plays an important role in RNA splicing (see p. 323).

Transcription Activators Can Promote Transcription by Releasing RNA Polymerase from Promoters

In some cases, transcription initiation requires that a DNA-bound transcription activator releases RNA polymerase from the promoter so as to allow it to begin transcribing the gene. In other cases, the RNA polymerase halts after transcribing about 50 nucleotides of RNA, and further elongation requires a transcription activator bound behind it (**Figure 7–21**). These paused polymerases are common in humans, where a significant fraction of genes that are not being transcribed have a paused polymerase located just downstream from the promoter.

The release of RNA polymerase can occur in several ways. In some cases, the activator brings in a chromatin remodeling complex that removes a nucleosome block to the elongating RNA polymerase. In other cases, the activator communicates with RNA polymerase (typically through a coactivator), signaling it to move ahead. Finally, as we saw in Chapter 6, RNA polymerase requires *elongation factors* to effectively transcribe through chromatin. In some cases, the key step in gene activation is the loading of these factors onto RNA polymerase, which can be directed by DNA-bound transcription activators. Once loaded, these factors allow the polymerase to move through blocks imposed by chromatin structure and begin transcribing the gene in earnest. Having RNA polymerase already poised on a promoter in the beginning stages of transcription bypasses the step of assembling many components at the promoter, which is often slow. This mechanism can therefore allow cells to begin transcribing a gene as a rapid response to an extracellular signal.

Transcription Activators Work Synergistically

We have seen that complexes of transcription activators and coactivators assemble cooperatively on DNA. We have also seen that these assemblies can promote 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

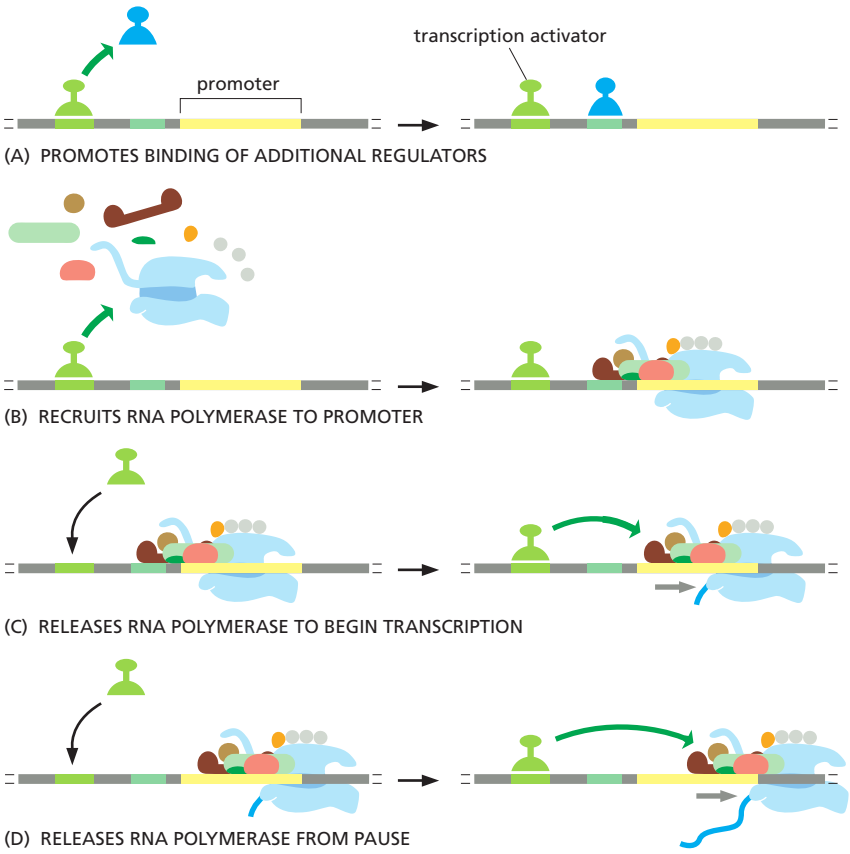


Figure 7–21 Transcription activators can act at different steps. In addition to (A) promoting binding of additional transcription regulators and (B) assembling RNA polymerase at promoters, transcription activators are often needed (C) to release already assembled RNA polymerases from promoters or (D) to release RNA polymerase molecules that become stalled after transcribing about 50 nucleotides of RNA. The activities shown in Figure 7–19 can affect each of these four steps.

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, transcription activators often exhibit *transcriptional synergy*, where several DNA-bound activator proteins working together produce a transcription rate that is much higher than the sum of their transcription rates working alone (**Figure 7–22**).

An important point is that a transcription activator protein must be bound to DNA 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, along with the coactivator proteins they bring to DNA.

Eukaryotic Transcription Repressors Can Inhibit Transcription in Several Ways

Although the "default" state of eukaryotic DNA packaged into nucleosomes is resistant to transcription, eukaryotes nonetheless use transcription regulators to

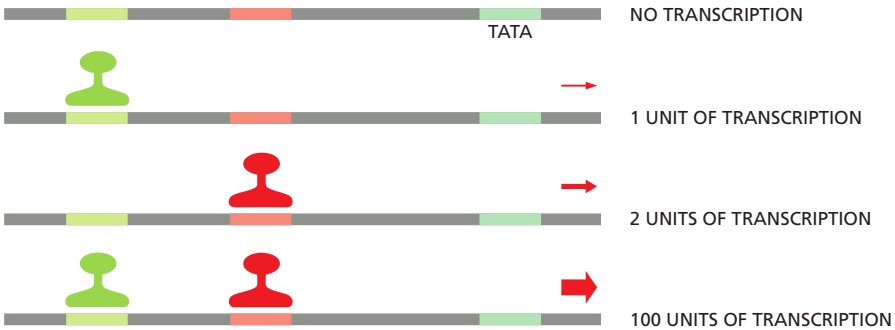


Figure 7–22 Transcriptional synergy. This experiment compares the rate of transcription produced by three experimentally constructed regulatory regions in a eukaryotic cell and reveals transcriptional synergy, a greater than additive effect of multiple activators working together. For simplicity, coactivators have been omitted from the diagram.

Such transcriptional synergy is not only observed between different transcription activators from the same organism; it is also seen between activator proteins from different eukaryotic species when they are experimentally introduced into the same cell. This last observation reflects the high degree of conservation of the machinery responsible for eukaryotic transcription initiation.

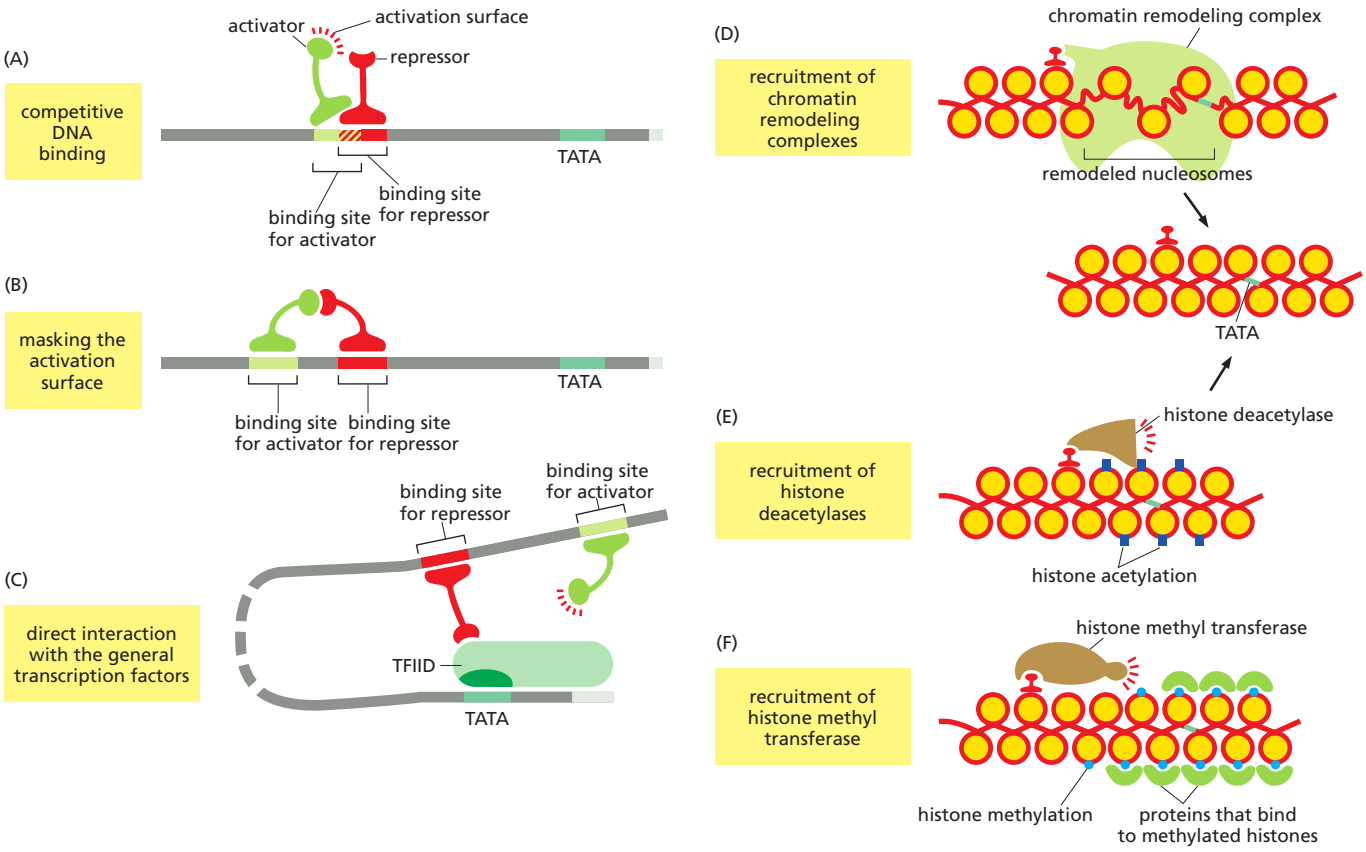


Figure 7–23 Six ways in which eukaryotic repressor proteins can operate. (A) Activator proteins and repressor proteins compete for binding to the same regulatory DNA sequence. (B) Both proteins bind DNA, but the repressor prevents the activator from carrying out its functions. (C) The repressor blocks assembly of 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 (see Figure 7–20), and the repressor simply reverses this modification. (F) The repressor attracts a histone methyl transferase, which modifies certain positions on histones by attaching methyl groups; the methylated histones, in turn, are bound by proteins that maintain the chromatin in a transcriptionally silent form.

repress the transcription of genes. These transcription repressors can both depress the rate of transcription below the default value and rapidly shut off genes that were previously activated. We saw in Chapter 4 that large regions of the genome can be shut down by the packaging of DNA into especially resistant forms of chromatin. However, eukaryotic genes are rarely organized along the genome according to function, and this strategy is not generally applicable for shutting off a set of genes that work together. Instead, most eukaryotic repressors work on a gene-by-gene basis. Unlike bacterial repressors, eukaryotic 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 7–23**. Although all of these mechanisms ultimately block transcription by RNA polymerase, eukaryotic transcription repressors typically act by bringing co-repressors to DNA. Like transcription activation, transcription repression can act through more than one mechanism at a given target gene, thereby ensuring especially 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.

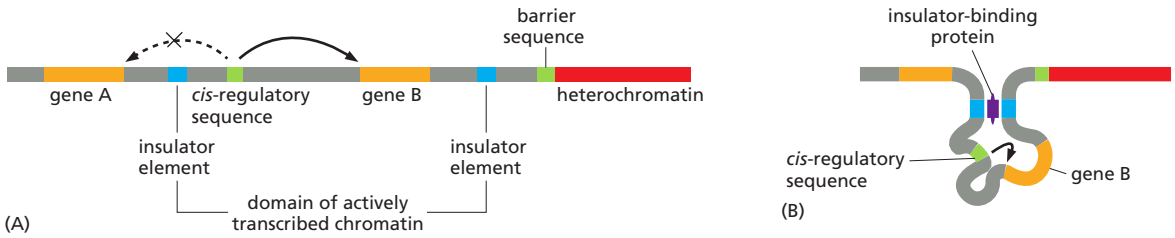


Figure 7–24 Schematic diagram summarizing the properties of insulators and barrier sequences. (A) Insulators directionally block the action of *cis*-regulatory sequences, whereas barrier sequences prevent the spread of heterochromatin. How barrier sequences likely function is depicted in Figure 4–41. (B) Insulator-binding proteins *(purple)* hold chromatin in loops, thereby favoring "correct" *cis*-regulatory sequence–gene associations. Thus, gene B is properly regulated, and gene B's *cis*-regulatory sequences are prevented from influencing the transcription of gene A.

Insulator DNA Sequences Prevent Eukaryotic Transcription Regulators from Influencing Distant Genes

We have seen that 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 eukaryotic transcription regulators can act across very long stretches of DNA, with the intervening DNA looped out. How, then, are control regions of different genes kept from interfering with one another? For example, what keeps a transcription regulator bound on the control region of one gene from looping in the wrong direction and inappropriately influencing the transcription of an adjacent gene?

To avoid such cross-talk, several types of DNA elements 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 *cis*-regulatory sequences from running amok and activating inappropriate genes (**Figure 7–24**). Insulators function by forming loops of chromatin, an effect mediated by specialized proteins that bind them (see Figures 4–48 and 7–24B). The loops hold a gene and its control region in rough proximity and help to prevent the control region from "spilling over" to adjacent genes. Importantly, these loops can be in different in different cell types, depending on the particular proteins and chromatin structures that are present.

The distribution of insulators and barrier sequences in a genome is thought to divide it into independent domains of gene regulation and chromatin structure (see pp. 207–208). Aspects of this organization can be visualized by staining whole chromosomes for the specialized proteins that bind these DNA elements (**Figure 7–25**).

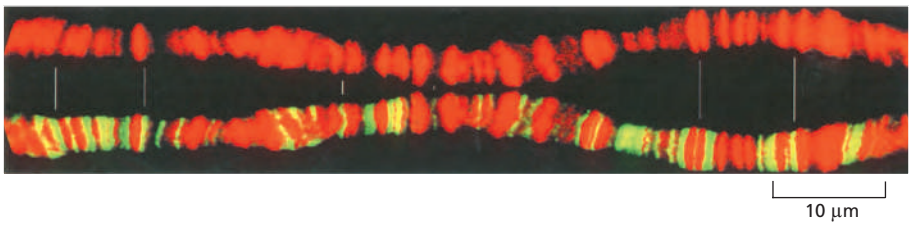


Figure 7–25 Localization of a *Drosophila* insulator-binding protein on polytene chromosomes. A polytene chromosome (discussed in Chapter 4) was stained with propidium iodide *(red)* to show its banding patterns, with bands appearing *bright red* and interbands as dark gaps in the pattern (top). The positions on this polytene chromosome that are bound by a particular insulator protein are stained *bright green* using antibodies directed against the protein (bottom). This protein 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.)

Although chromosomes are organized into orderly domains that discourage control regions from acting indiscriminately, there are special circumstances where a control region located on one chromosome has been found to activate a gene located on a different chromosome. Although there is much we do not understand about this mechanism, it indicates the extreme versatility of transcriptional regulation strategies.

Summary

Transcription regulators switch the transcription of individual genes on and off in cells. In prokaryotes, these proteins typically 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 eukaryotic genes is much more complex, commensurate with a larger genome size and the large variety of cell types that are formed. A single eukaryotic gene is typically controlled by many transcription regulators bound to sequences that can be tens or even hundreds of thousands of nucleotide pairs from the promoter that directs transcription of the gene. Eukaryotic activators and repressors act by a wide variety of mechanisms—generally altering chromatin structure and controlling the assembly of the general transcription factors and RNA polymerase at the promoter. They do this by attracting coactivators and co-repressors, protein complexes that perform the necessary biochemical reactions. The time and place that each gene is transcribed, as well as its rates of transcription under different conditions, are determined by the particular spectrum of transcription regulators that bind to the regulatory region of the gene.

MOLECULAR GENETIC MECHANISMS THAT CREATE AND MAINTAIN 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. 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 eukaryotes, as well as most such changes 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.

In this section, we shall examine not only cell memory mechanisms, but also how gene regulatory devices can be combined to create the "logic circuits" through which cells integrate signals and remember events in their past. We begin by considering one such complex gene control region in detail.

Complex Genetic Switches That Regulate *Drosophila* Development Are Built Up from Smaller Molecules

We have seen that transcription regulators can be positioned at multiple sites along long stretches of DNA and that these proteins can bring into play coactivators and co-repressors. Here, we discuss how the numerous transcription regulators that are bound to the control region of a gene can cause the gene to be transcribed at the proper place and time.

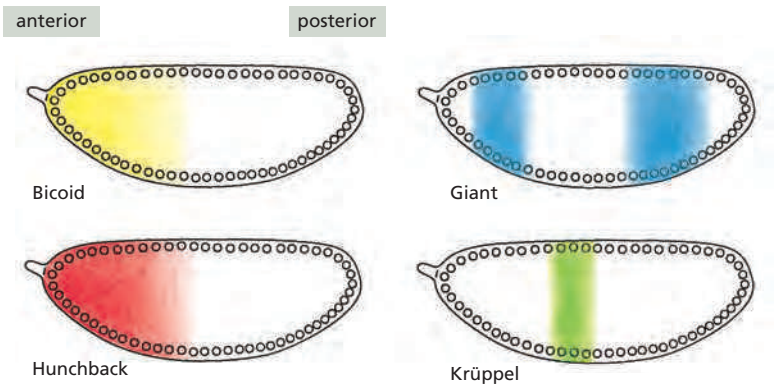


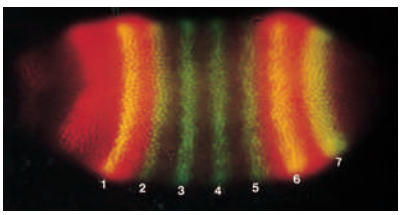
Figure 7–26 The nonuniform distribution of transcription regulators in an early *Drosophila* embryo. At this stage, the embryo is a syncytium; that is, multiple nuclei are contained in a common cytoplasm. Although not shown in these drawings, all of these proteins are concentrated in the nuclei. How such differences are established is discussed in Chapter 21.

Consider 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 21, 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 contains a mixture of transcription regulators that are distributed unevenly along the length of the embryo, thus providing *positional information* that distinguishes one part of the embryo from another (**Figure 7–26**). Although the nuclei are initially identical, they rapidly begin to express different genes because they are exposed to different transcription regulators. For example, the nuclei near the anterior end of the developing embryo are exposed to a set of transcription regulators that is distinct from the set that influences nuclei at the middle or at the posterior end of the embryo.

The regulatory DNA sequences that control the *Eve* gene have evolved to "read" the concentrations of transcription regulators at each position along the length of the embryo, and they cause the *Eve* gene to be expressed in seven precisely positioned stripes, each initially five to six nuclei wide (**Figure 7–27**). How is this remarkable feat of information processing carried out? Although there is still much to learn, several general principles have emerged from studies of *Eve* and other genes that are similarly regulated.

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 *cis*-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 was 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, 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 7–28**). Similar experiments reveal the existence of other regulatory modules, each of which specifies other stripes.

Figure 7–27 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 pattern for the Giant protein is also shown in Figure 7–26. (Courtesy of Michael Levine.)



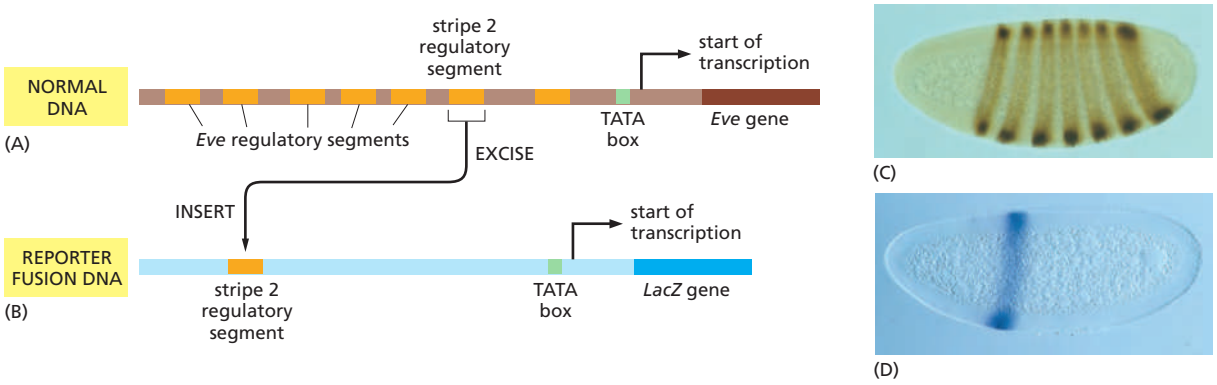


Figure 7–28 Experiment demonstrating the modular construction of the *Eve* gene regulatory region. (A) A 480-nucleotide-pair section of the *Eve* regulatory region was removed and (B) inserted upstream of a test promoter that directs the synthesis of the enzyme β-galactosidase (the product of the *E. coli LacZ* gene—see Figure 7–15). (C, D) When this artificial construct was reintroduced into the genome of *Drosophila* embryos, the embryos (D) expressed β-galactosidase (detectable by histochemical staining) precisely in the position of the second of the seven *Eve* stripes (C). β-Galactosidase is simple to detect and thus provides a convenient way to monitor the expression specified by a gene control region. As used here, β-galactosidase is said to serve as a reporter, since it "reports" the activity of a gene control region. (C and D, courtesy of Stephen Small and Michael Levine.)

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. The module contains recognition sequences for two transcription regulators (Bicoid and Hunchback) that activate *Eve* transcription and for two transcription regulators (Krüppel and Giant) that repress it (**Figure 7–29**). 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 7–30** shows the distributions of the four transcription regulators 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 stripe 2 module, whereas both Bicoid and Hunchback must bind for this module's maximal activation. This simple regulatory scheme suffices to turn on the stripe 2 module (and therefore the expression of the *Eve* gene) only in those nuclei located where the levels of both Bicoid and Hunchback are high and both Krüppel and Giant are absent—a combination that occurs in only one region of the early embryo. It is not known exactly how these four transcription regulators interact with coactivators and co-repressors to specify the final level of transcription across the stripe, but the outcome very likely relies on competition between activators and repressors that act by the mechanisms outlined in Figures 7–17, 7–19, and 7–23.

The stripe 2 element is autonomous, inasmuch as it specifies stripe 2 when isolated from its normal context (see Figure 7–28). The other stripe regulatory modules are thought to be constructed similarly, reading positional information provided by other combinations of transcription regulators. The entire *Eve* gene control region binds more than 20 different transcription regulators. Seven combinations of regulators—one combination for each stripe—specify *Eve* expression, while many other combinations (all those found in the interstripe regions of

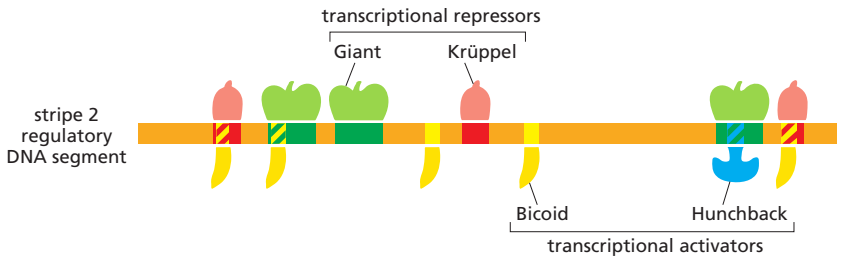


Figure 7–29 The *Eve* stripe 2 unit. The segment of the *Eve* gene control region identified in Figure 7–28 contains *cis*-regulatory sequences for four transcription regulators. It is known from genetic experiments that these four regulatory proteins are responsible for the proper expression of *Eve* in stripe 2. Flies that are deficient in the two gene activators Bicoid and Hunchback, for example, fail to efficiently express *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. As indicated, in some cases the binding sites for the transcription regulators 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 is mutually exclusive.

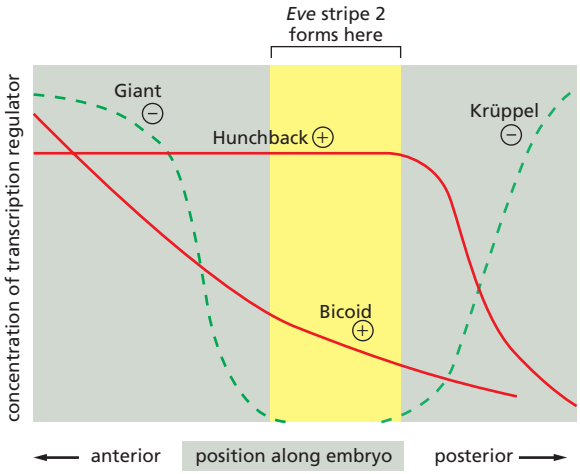


Figure 7–30 Distribution of the transcription regulators responsible for ensuring that *Eve* is expressed in stripe 2. The distributions of these proteins were visualized by staining a developing *Drosophila* embryo with antibodies directed against each of the four proteins. The expression of *Eve* in 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 are inactivated by mutation (see also Figures 7–26 and 7–27).

the embryo) keep the stripe elements silent. 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 *cis*-regulatory sequences recognized by specific transcription regulators.

The *Eve* gene itself encodes a transcription regulator, which, after its pattern of expression is set up in seven stripes, controls 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 21.

Eve exemplifies the complex control regions found in plants and animals. As this example shows, control regions can respond to many different inputs, integrate this information, and produce a complex spatial and temporal output as development proceeds. However, exactly how all these mechanisms work together to produce the final output is understood only in broad outline (**Figure 7–31**).

Transcription Regulators Are Brought Into Play by Extracellular Signals

The above example from *Drosophila* clearly illustrates the power of combinatorial control, but this case is unusual in that the nuclei are exposed directly to positional cues in the form of concentrations of transcription regulators. In embryos of most other organisms and in all adults, individual nuclei are in separate cells, and extracellular information (including positional cues) must be passed across the plasma membrane so as to generate signals in the cytosol that cause different transcription regulators to become active in different cell types. Some of the different mechanisms that are known to be used to activate transcription regulators are diagrammed in **Figure 7–32**, and in Chapter 15, we discuss how extracellular signals trigger these changes.

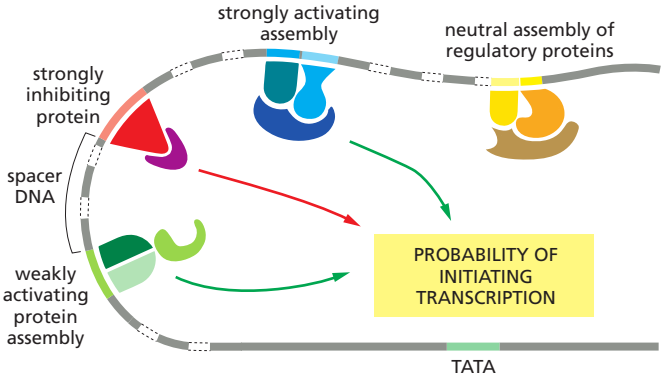


Figure 7–31 The integration of multiple inputs at a promoter. Multiple sets of transcription regulators, coactivators, and co-repressors can work together to influence transcription initiation at a promoter, as they do in the *Eve* stripe 2 module illustrated in Figure 7–29. It is not yet understood 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 7–17, 7–19, and 7–23.

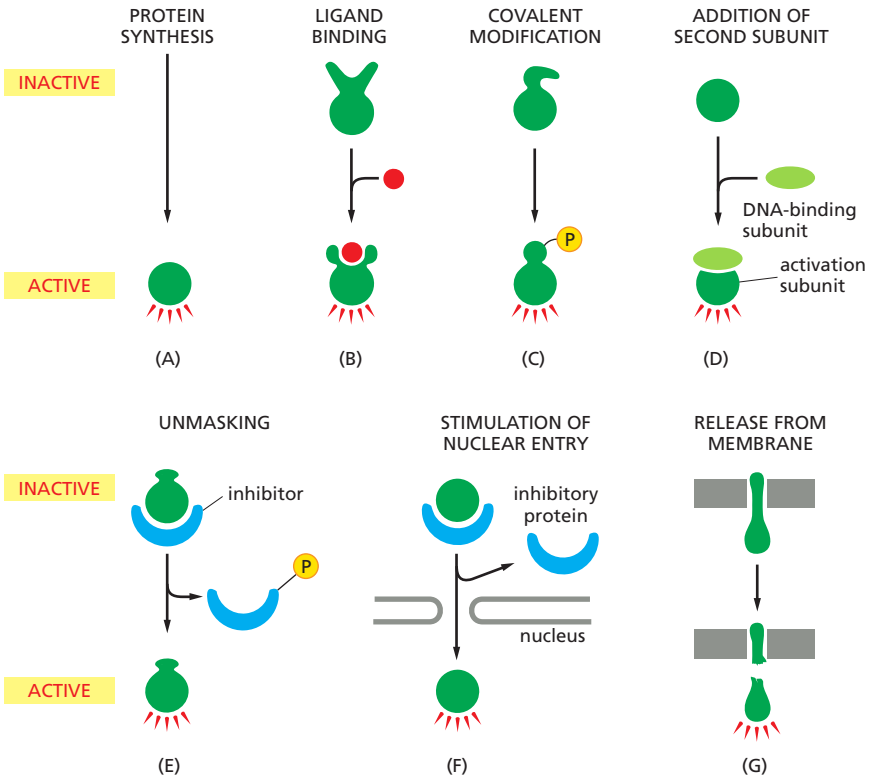


Figure 7–32 Some ways in which the activity of transcription regulators is controlled inside eukaryotic cells. (A) The protein is synthesized only when needed and is rapidly degraded by proteolysis so that it does not accumulate. (B) Activation by ligand binding. (C) Activation by covalent modification. Phosphorylation is shown here, but many other modifications are possible (see Table 3–3, p. 165). (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 transcription regulator from a membrane bilayer by regulated proteolysis.

Combinatorial Gene Control Creates Many Different Cell Types

We have seen that transcription regulators can act in combination to control the expression of an individual gene. It is also generally true that each transcription regulator in an organism contributes to the control of many genes. This point is illustrated schematically in **Figure 7–33**, which shows how combinatorial gene control makes it possible to generate a great deal of biological complexity even with relatively few transcription regulators.

Due to combinatorial control, a given transcription regulator 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, transcription regulators 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.

Combinatorial gene control causes the effect of adding a new transcription regulator to a cell to depend on that cell's past history, since it is this history that determines which transcription regulators are already present. Thus, during development, a cell can accumulate a series of transcription regulators that need not initially alter gene expression. The addition of the final members of the requisite combination of transcription regulators will complete the regulatory message, and can lead to large changes in gene expression.

The importance of combinations of transcription regulators for the specification of cell types is most easily demonstrated by their ability—when expressed artificially—to convert one type of cell to another. Thus, the artificial expression of three neuron-specific transcription regulators in liver cells can convert the liver cells into functional nerve cells (**Figure 7–34**). In some cases, expression of even a single transcription regulator is sufficient to convert one cell type to another. For example, when the gene encoding the transcription regulator MyoD is artificially introduced into fibroblasts cultured from skin connective tissue, the fibroblasts form muscle-like cells. As discussed in Chapter 22, fibroblasts, which are derived from the same broad class of embryonic cells as muscle cells, have already accumulated many of the other necessary transcription regulators required for the

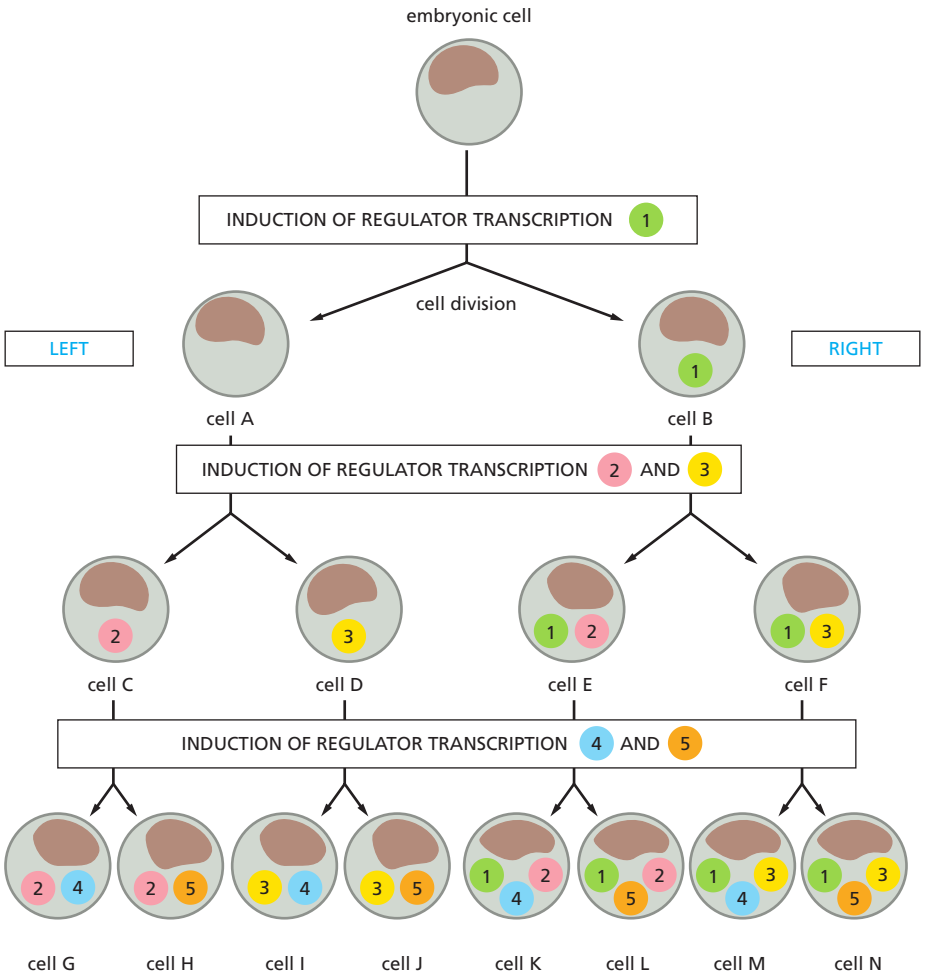


Figure 7–33 The importance of combinatorial gene control for development. Combinations of a few transcription regulators can generate many cell types during development. In this simple, idealized scheme, a "decision" to make one of a pair of different transcription regulators (shown as numbered circles) 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 transcription regulator is assumed to be self-perpetuating once it has become initiated (see Figure 7–39). In this way, through cell memory, the final combinatorial specification is built up step by step. In this purely hypothetical example, five different transcription regulators have created eight final cell types (G–N).

combinatorial control of the muscle-specific genes, and the addition of MyoD completes the unique combination required to direct the cells to become muscle. An even more striking example is seen by artificially expressing, early in development, a single *Drosophila* transcription regulator (Eyeless) in groups of cells

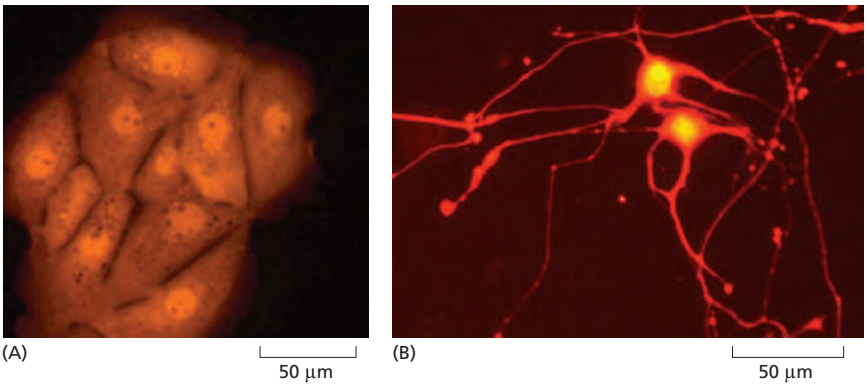


Figure 7–34 A small set of transcription regulators can convert one differentiated cell type into another. In this experiment, (A) liver cells grown in culture were converted into (B) neuronal cells via the artificial expression of three nerve-specific transcription regulators. Both types of cells express an artificial *red* fluorescent protein, which is used to visualize them. This conversion involves the activation of many nerve-specific genes as well as the repression of many liver-specific genes. (From S. Marro et al., *Cell Stem Cell* 9:374–382, 2011. With permission from Elsevier.)

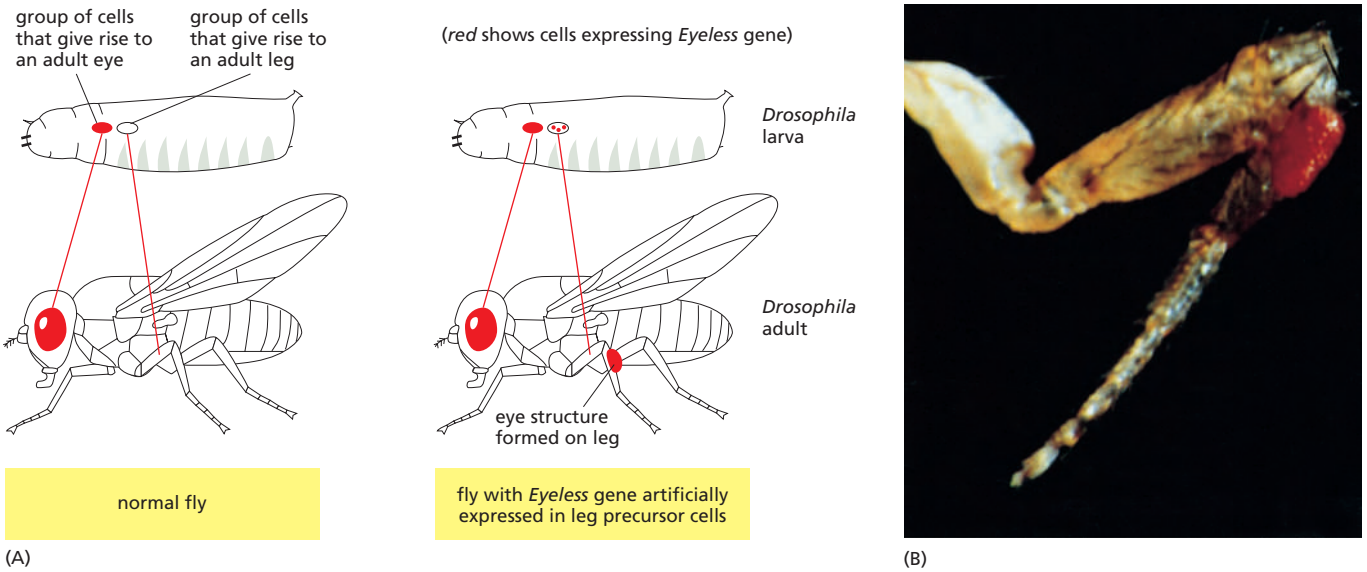


Figure 7–35 Expression of the *Drosophila Eyeless* gene in precursor cells of the leg triggers the development of an eye on the leg. (A) Simplified diagrams showing the result when a fruit fly larva contains either the normally expressed *Eyeless* gene *(left)* or an *Eyeless* gene that is additionally 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 21–2). The transcription regulator was named Eyeless because its inactivation in otherwise normal flies causes the loss of eyes. (B, courtesy of Walter Gehring.)

that would normally go on to form leg parts. Here, this abnormal gene expression change causes eye-like structures to develop in the legs (**Figure 7–35**).

Specialized Cell Types Can Be Experimentally Reprogrammed to Become Pluripotent Stem Cells

Manipulation of transcription regulators can also coax various differentiated cells to *de-differentiate* into pluripotent stem cells that are capable of giving rise to the different cell types in the body, much like the embryonic stem (ES) cells discussed in Chapter 22. When three specific transcription regulators are artificially expressed in cultured mouse fibroblasts, a number of cells become **induced pluripotent stem cells (iPS cells)**—cells that look and behave like the pluripotent ES cells that are derived from embryos (**Figure 7–36**). This approach has been adapted to produce iPS cells from a variety of specialized cell types, including cells taken from humans. Such human iPS cells can then be directed to generate a population of differentiated cells for use in the study or treatment of disease, as we discuss in Chapter 22.

Although it was once thought that cell differentiation was irreversible, it is now clear that by manipulating combinations of **master transcription regulators**, cell types and differentiation pathways can be readily altered.

Combinations of Master Transcription Regulators Specify Cell Types by Controlling the Expression of Many Genes

As we saw in the introduction of this chapter, different cell types of multicellular organisms differ enormously in the proteins and RNAs they express. For example, only muscle cells express special types of actin and myosin that form the contractile

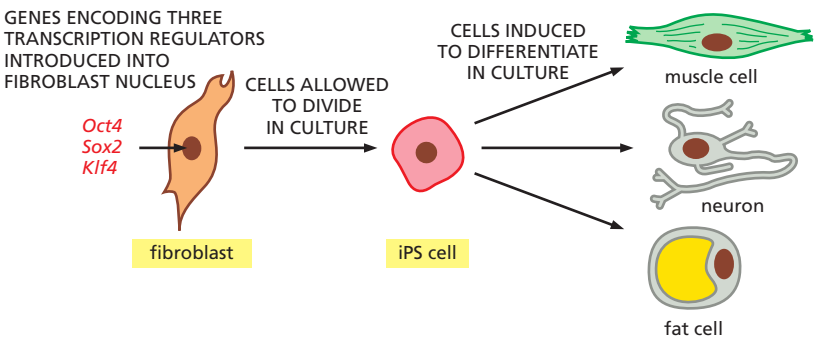


Figure 7–36 A combination of transcription regulators can induce a differentiated cell to de-differentiate into a pluripotent cell. The artificial expression of a set of three genes, each of which encodes a transcription regulator, can reprogram a fibroblast into a pluripotent cell with embryonic stem (ES)-cell-like properties. Like ES cells, such induced pluripotent stem (iPS) cells can proliferate indefinitely in culture and can be stimulated by appropriate extracellular signal molecules to differentiate into almost any cell type found in the body. Transcription regulators such as Oct4, Sox2, and Klf4 are often called *master transcription regulators* because their expression is sufficient to trigger a change in cell identity.

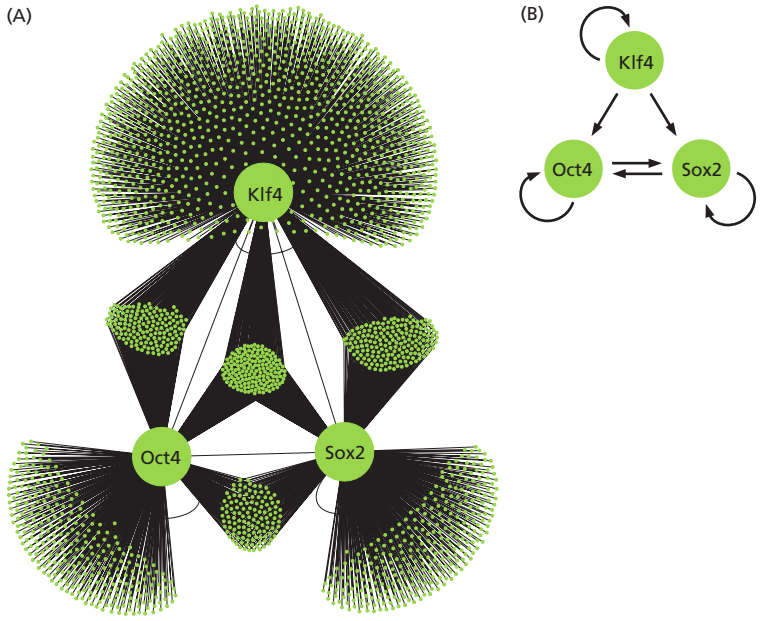


Figure 7–37 A portion of the transcription network specifying embryonic stem cells. (A) The three master transcription regulators in Figure 7–36 are shown as large circles. Genes whose *cis*-regulatory sequences are bound by each regulator in embryonic stem cells are indicated by a small dot (representing the gene) connected by a thin line (representing the binding reaction). Note that many of the target genes are bound by more than one of the regulators. (B) The master regulators control their own expression. As shown here, the three transcriptional regulators bind to their own control regions (indicated by feedback loops), as well as those of the other master regulators (indicated by straight arrows). (Courtesy of Trevor Sorrells, based on data from J. Kim et al., *Cell* 132:1049–1061, 2008.)

apparatus, while nerve cells must make and assemble all the proteins needed to form dendrites and synapses. We have seen that these patterns of cell-type-specific expression are orchestrated by a combination of master transcription regulators. In many cases, these proteins bind directly to *cis*-regulatory sequences of the genes particular to that cell type. Thus, MyoD binds directly to *cis*-regulatory sequences located in the control regions of the muscle-specific genes. In other cases, the master regulators control the expression of "downstream" transcription regulators which, in turn, bind to the control regions of other cell-type-specific genes and control their synthesis.

The specification of a particular cell type typically involves changes in the expression of several thousand genes. Genes whose protein products are required in the cell type are expressed at high levels, while those not needed are typically down-regulated. As might be imagined, the pattern of binding between the master regulators and all of the regulated genes can be extremely elaborate (**Figure 7–37**). When we consider that many of these regulated genes have control regions that span tens of thousands of nucleotide pairs, commensurate with the *Eve* example discussed above, we can begin to appreciate the enormous complexity of cell-type specification.

An outstanding question in biology is how the information in a genome is used to specify a multicellular organism. Although we have the general outline of the answer, we are far from understanding how a single cell type is completely specified, let alone a whole organism.

Specialized Cells Must Rapidly Turn Sets of Genes On and Off

Although they generally maintain their identities, specialized cells must constantly respond to changes in their environment. Among the most important changes are signals from other cells that coordinate the behavior of the whole organism. Many of these signals induce transient changes in gene transcription, and we discuss the nature of these signals in detail in Chapter 15. Here, we consider how specialized cell types rapidly and decisively switch groups of genes on and off in response to their environment. Even though control of gene expression is combinatorial, the effect of a single transcription regulator 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. Likewise, the addition of a particular protein can turn on many different genes.

An example is the rapid control of gene expression by the human glucocorticoid receptor protein. To bind to its *cis*-regulatory sequences in the genome, this transcription regulator must first form a complex with a molecule of a glucocorticoid steroid hormone, such as cortisol (see Figure 15–64). 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, such as tyrosine aminotransferase, as we discussed earlier in this chapter (see Figure 7–3). Although these genes all have different and complex control regions, their maximal expression depends on the binding of the hormone–glucocorticoid receptor complex to its *cis*-regulatory sequence, which is present in the control region 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 transcription regulator can rapidly control the expression of many different genes (**Figure 7–38**).

The effects of the glucocorticoid receptor are not confined to cells of the liver. In other cell types, activation of this transcription regulator by hormone also causes changes in the expression levels of many genes; the genes affected, however, are usually different from those affected in liver cells. As we have seen, each cell type has an individualized set of transcription regulators, 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 transcription regulators, switching it on with hormone produces a different spectrum of effects in each cell type.

Differentiated Cells Maintain Their Identity

Once a cell has become differentiated into a particular cell type, it will generally remain differentiated, and all its progeny cells will remain that same cell type. Some highly specialized cells, including skeletal muscle cells and neurons, never divide again once they have differentiated—that is, they are *terminally differentiated* (as discussed in Chapter 17). But many other differentiated cells—such as

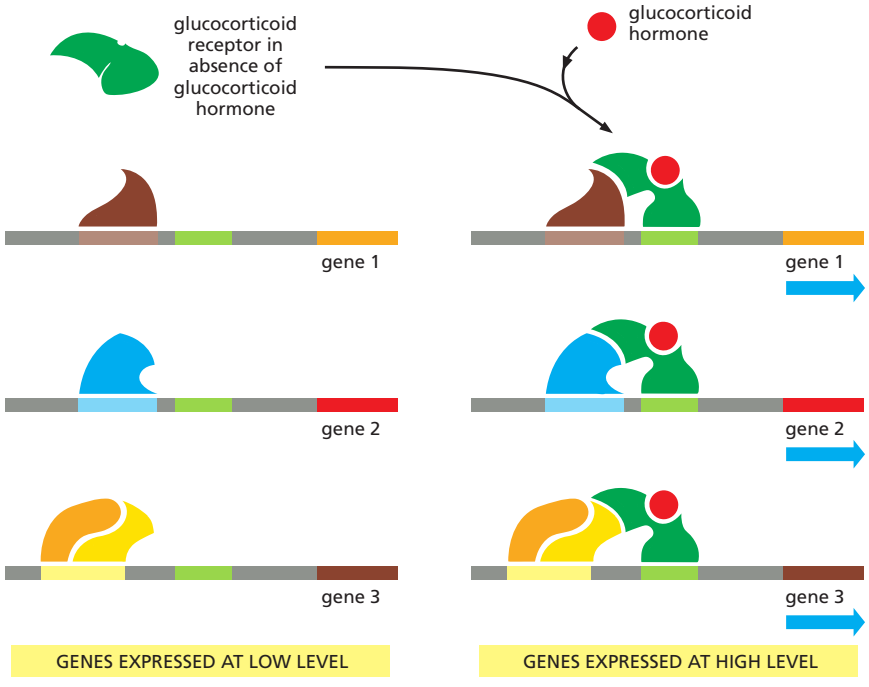


Figure 7–38 A single transcription regulator can coordinate the expression of many different genes. The action of the glucocorticoid receptor is illustrated schematically. On the *left* is a series of genes, each of which has various transcription regulators bound to its regulatory region. However, these bound proteins are not sufficient on their own to fully activate transcription. On the *right* is shown the effect of adding an additional transcription regulator—the glucocorticoid receptor in a complex with glucocorticoid hormone—that has a *cis*-regulatory sequence in the control region of each gene. The glucocorticoid receptor completes the combination of transcription regulators required for maximal initiation of transcription, and the genes are now switched on as a set. When the hormone is no longer present, the glucocorticoid receptor dissociates from DNA and the genes return to their pre-stimulated levels.

fibroblasts, smooth muscle cells, and liver cells—will divide many times in the life of an individual. When they do, these specialized cell types give rise only to cells like themselves: smooth muscle cells do not give rise to liver cells, nor liver cells to fibroblasts.

For a proliferating cell to maintain its identity—a property called **cell memory**—the patterns of gene expression responsible for that identity must be remembered and passed on to its daughter cells through subsequent cell divisions. Thus, in the model we discussed in Figure 7–33, the production of each transcription regulator, once begun, has to be continued in the daughter cells of each cell division. How is such perpetuation accomplished?

Cells have several ways of ensuring that their daughters "remember" what kind of cells they are. One of the simplest and most important is through a positive feedback loop, where a master cell-type transcription regulator activates transcription of its own gene, in addition to that of other cell-type-specific genes. Each time a cell divides, the regulator is distributed to both daughter cells, where it continues to stimulate the positive feedback loop, making more of itself each division. Positive feedback is crucial for establishing "self-sustaining" circuits of gene expression that allow a cell to commit to a particular fate—and then to transmit that information to its progeny (**Figure 7–39**).

As was previously shown in Figure 7–37B, the master regulators needed to maintain the pluripotency of iPS cells bind to *cis*-regulatory sequences in their own control regions, providing examples of the type of positive feedback loop. In addition, most of these pluripotent cell regulators also activate transcription of other master regulators, resulting in a complex series of indirect feedback loops. For example, if A activates B, and B activates A, this forms a positive feedback loop where A activates its own expression, albeit indirectly. The series of direct and indirect feedback loops observed in the iPS circuit is typical of other specialized cell circuits. Such a network structure strengthens cell memory, increasing the probability that a particular pattern of gene expression is transmitted through successive generations. For example, if the level of A drops below the critical threshold to stimulate its own synthesis, regulator B can rescue it. By successive application of this mechanism, a complex series of positive feedback loops among multiple transcription regulators can stably maintain a differentiated state through many cell divisions.

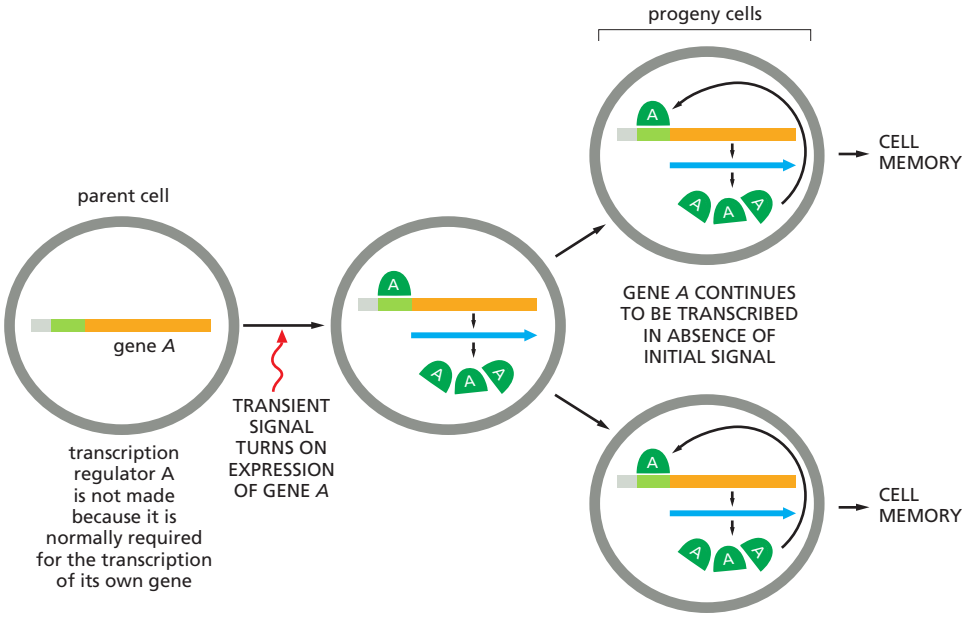


Figure 7–39 A positive feedback loop can create cell memory. Protein A is a master transcription regulator that activates the transcription of its own gene—as well as other cell-type-specific genes (not shown). All of the descendants of the original cell will therefore "remember" that the progenitor cell had experienced a transient signal that initiated the production of protein A.

Positive feedback loops formed by transcription regulators are probably the most prevalent way of ensuring that daughter cells remember what kind of cells they are meant to be, and they are found in all species on Earth. For example, many bacteria and single-cell eukaryotes form different types of cells, and positive feedback loops lie at the heart of mechanisms that maintain their cell types through many rounds of cell division. Plants and animals also make extensive use of transcription feedback loops; as we shall discuss later in the chapter, they have additional, more specialized mechanisms for making cell memory even stronger. But first, we briefly consider how combinations of transcription regulators and *cis*-regulatory sequences can be combined to create useful logic devices for the cell.

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 different types of operations. An analysis of gene regulatory circuits reveals that certain simple types of arrangements (called *network motifs*) are found over and over again in cells from widely different species. For example, positive and negative feedback loops are common in all cells (**Figure 7–40**). Whereas the former provides a simple memory device, the latter is often used to keep the expression of a gene close to a standard level despite the variations in biochemical conditions inside a cell. Suppose, for example, that a transcription repressor protein binds to the regulatory region of its own gene and exerts a strong negative feedback, such that transcription falls to a very low rate when the concentration of the repressor protein is above some critical value (determined by its affinity for its DNA binding site). The concentration of the protein can then be held close to the critical value, since any circumstance that causes a fall below that value can lead to a steep increase in synthesis, and any that causes a rise above that value will lead to synthesis being switched 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. If there is a delay in the feedback loop, the result may be spontaneous oscillations in the expression of the gene (see Figure 15–18). The different types of behavior produced by a feedback loop will depend on the details of the system; for example, how tightly the transcription regulator binds to its *cis*-regulatory sequence, its rate of synthesis, and its rate of decay. We discuss these issues in quantitative terms and in more detail in Chapter 8.

With two or more transcription regulators, the possible range of circuit behaviors becomes more complex. Some bacterial viruses contain 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; such a loop can serve as a filter, responding to input signals that are prolonged but disregarding those that are brief (**Figure 7–41**). These various network motifs resemble miniature logic devices, and they can process information in surprisingly sophisticated ways.

The simple types of devices just illustrated are found to be interwoven in eukaryotic cells, creating exceedingly complex circuits (**Figure 7–42**). Each cell in a developing multicellular organism is equipped with similarly complex control

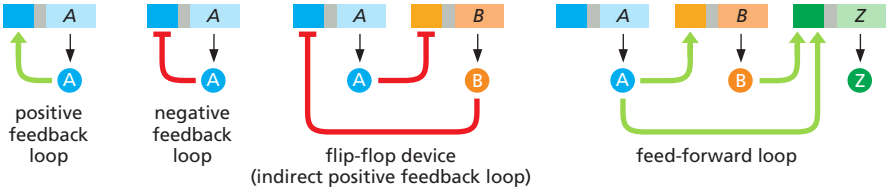


Figure 7–40 Common types of network motifs in transcription circuits. A and B represent transcription regulators, arrows indicate positive transcription control, while lines with bars depict negative transcription control. In the feed-forward loop, A and B represent transcription regulators that both activate the transcription of target gene Z (see also Figure 8–86).

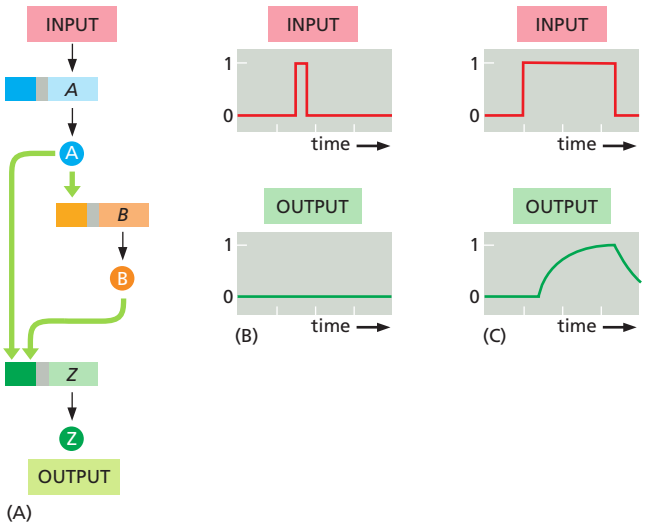


Figure 7–41 How a feed-forward loop can measure the duration of a signal. (A) In this theoretical example, transcription regulators A and B are both required for transcription of *Z*, and A becomes active only when an input signal is present. (B) If the input signal to A is brief, A does not stay active 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 the product of Z, along with 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 discern—and subsequently test—the different ways in which they may function (see Figure 8–86). (Adapted from S.S. Shen-Orr et al., *Nat. Genet.* 31:64–68, 2002. With permission from Macmillan Publishers Ltd.)

machinery, and it must, in effect, use its intricate system of interlocking transcription 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 how to study such complex intracellular control networks. Indeed, without new approaches, coupled with quantitative information that is far more precise and complete than we now possess, it will be impossible to predict the behavior of a system such as that shown in Figure 7–42. As explained in Chapter 8, a circuit diagram by itself is not enough.

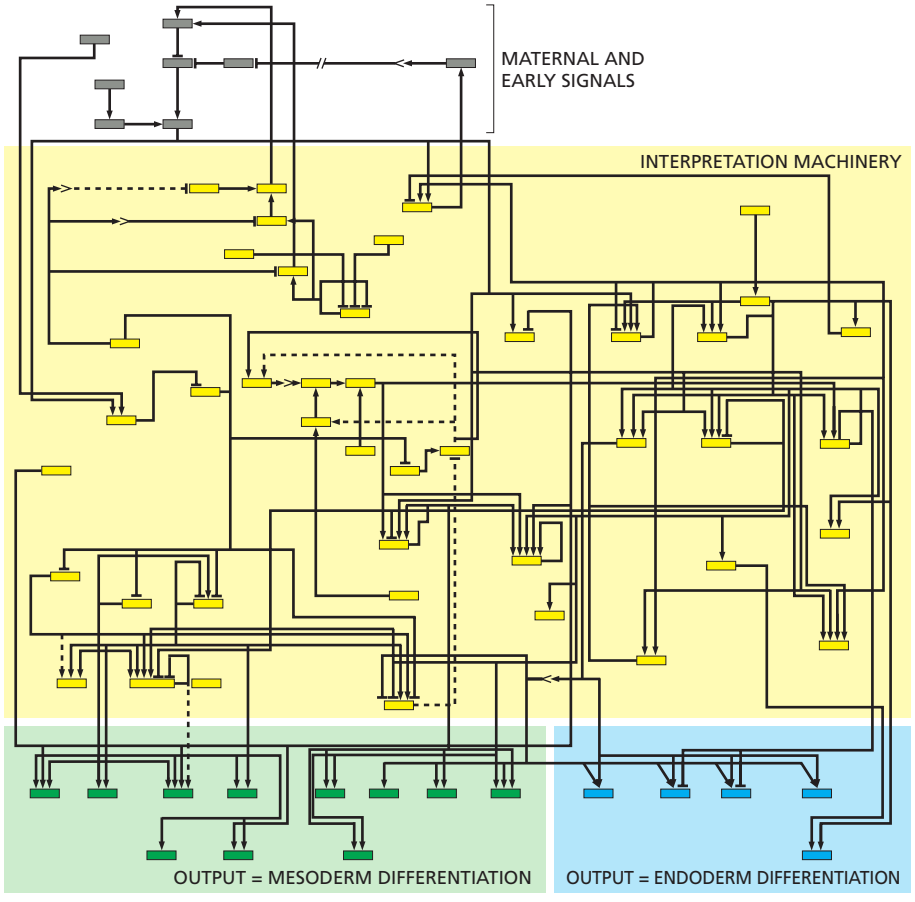


Figure 7–42 The exceedingly complex gene circuit that specifies a portion of the developing sea urchin embryo. Each *colored small box* represents a different gene. Those in *yellow* code for transcription regulators and those in *green* and *blue* code for proteins that give cells of the mesoderm and 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. As in Figure 7–40, arrows depict instances in which a transcription regulator activates the transcription of another gene. Lines ending in bars indicate examples of gene repression. (From I.S. Peter and E.H. Davidson, *Nature* 474:635–639, 2011. With permission from Macmillan Publishers Ltd.)

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. The transcription of any particular gene is generally controlled by a combination of transcription regulators. Each type of cell in a higher eukaryotic organism contains a specific set of transcription regulators that ensures the expression of only those genes appropriate to that type of cell. A given transcription regulator may be active in a variety of circumstances and is typically involved in the regulation of many different genes.

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. Direct or indirect positive feedback loops, which enable transcription regulators to perpetuate their own synthesis, provide the simplest mechanism for cell memory. Transcription circuits also provide the cell with the means to carry out other types of logic operations. Simple transcription circuits combined into large regulatory networks drive highly sophisticated programs of embryonic development that will require new approaches to decipher.

MECHANISMS THAT REINFORCE CELL MEMORY IN PLANTS AND ANIMALS

Thus far in this chapter, we have emphasized the regulation of gene transcription by proteins that associate either directly or indirectly with DNA. However, DNA itself can be covalently modified, and certain types of chromatin states appear to be inherited. In this section, we shall see how these phenomena also provide opportunities for the regulation of gene expression. At the end of this section, we discuss how, in mice and humans, an entire chromosome can be transcriptionally inactivated using such mechanisms, and how this state can be maintained through many cell divisions.

Patterns of DNA Methylation Can Be Inherited When Vertebrate Cells Divide

In vertebrate cells, the methylation of cytosine provides a mechanism through which gene expression patterns can be passed on to progeny cells. The methylated form of cytosine, 5-methyl cytosine (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 7–43**). **DNA methylation** in vertebrate DNA occurs on cytosine (C) nucleotides largely 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 methyl transferase* 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 7–44**).

Although DNA methylation patterns can be maintained in differentiated cells by the mechanism shown in Figure 7–44, methylation patterns are dynamic during mammalian 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 methyl transferase activity, resulting in the passive loss of methyl groups during each round of DNA replication, or by *demethylating enzymes* (discussed below). Later in development, new methylation patterns are established by several *de novo DNA methyl transferases* that are directed to DNA by sequence

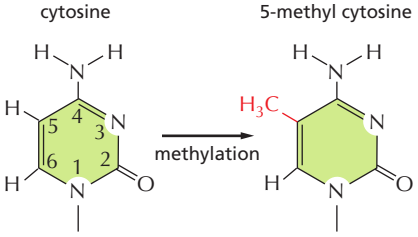


Figure 7–43 Formation of 5-methyl cytosine occurs by methylation of a cytosine base in the DNA double helix. In vertebrates, this event is largely confined to selected cytosine (C) nucleotides located in the sequence CG. CG sequences are sometimes denoted as CpG sequences, where the p indicates a phosphate linkage to distinguish it from a CG base pair. In this chapter, we will continue to use the simpler nomenclature CG to indicate this dinucleotide.

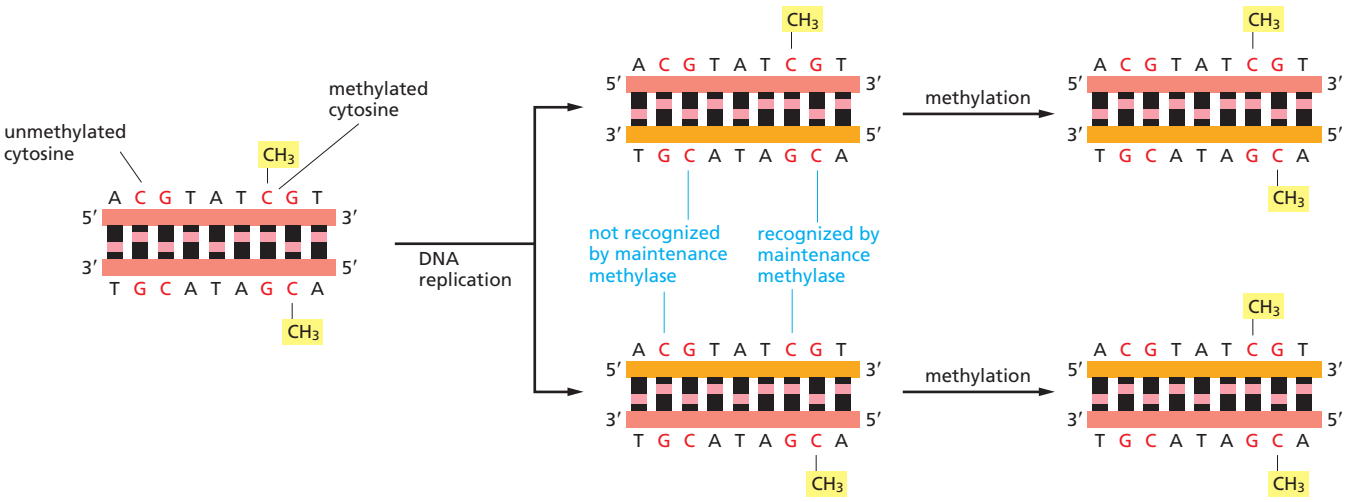


Figure 7–44 How DNA methylation patterns are faithfully inherited. In vertebrate DNA, a large fraction of the cytosine nucleotides in the sequence CG is methylated (see Figure 7–43). Because of the existence of a methyl-directed methylating enzyme (the maintenance methyl transferase), once a pattern of DNA methylation is established, that pattern of methylation is inherited in the progeny DNA, as shown.

specific DNA-binding proteins. 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. A very important role is to work in conjunction with other gene expression control mechanisms to establish a particularly efficient form of gene repression. This combination of mechanisms ensures that unneeded eukaryotic 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.

DNA methylation helps to repress transcription in several ways. The methyl groups on methylated cytosines lie in the major groove of DNA and interfere directly with the binding of proteins (transcription regulators as well as the general transcription factors) required for transcription initiation. In addition, the cell contains a repertoire of proteins that bind specifically to methylated DNA.The best characterized of these associate with histone modifying enzymes, leading to a repressive chromatin state where chromatin structure and DNA methylation act synergistically (**Figure 7–45**). One reflection of the importance of DNA methylation to humans is the widespread involvement of "incorrect" DNA methylation patterns in cancer progression (discussed in Chapter 20).

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 vertebrate genome tend to be eliminated in the course of evolution. Accidental deamination of an unmethylated C gives rise to U (see Figure 5–38), 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.

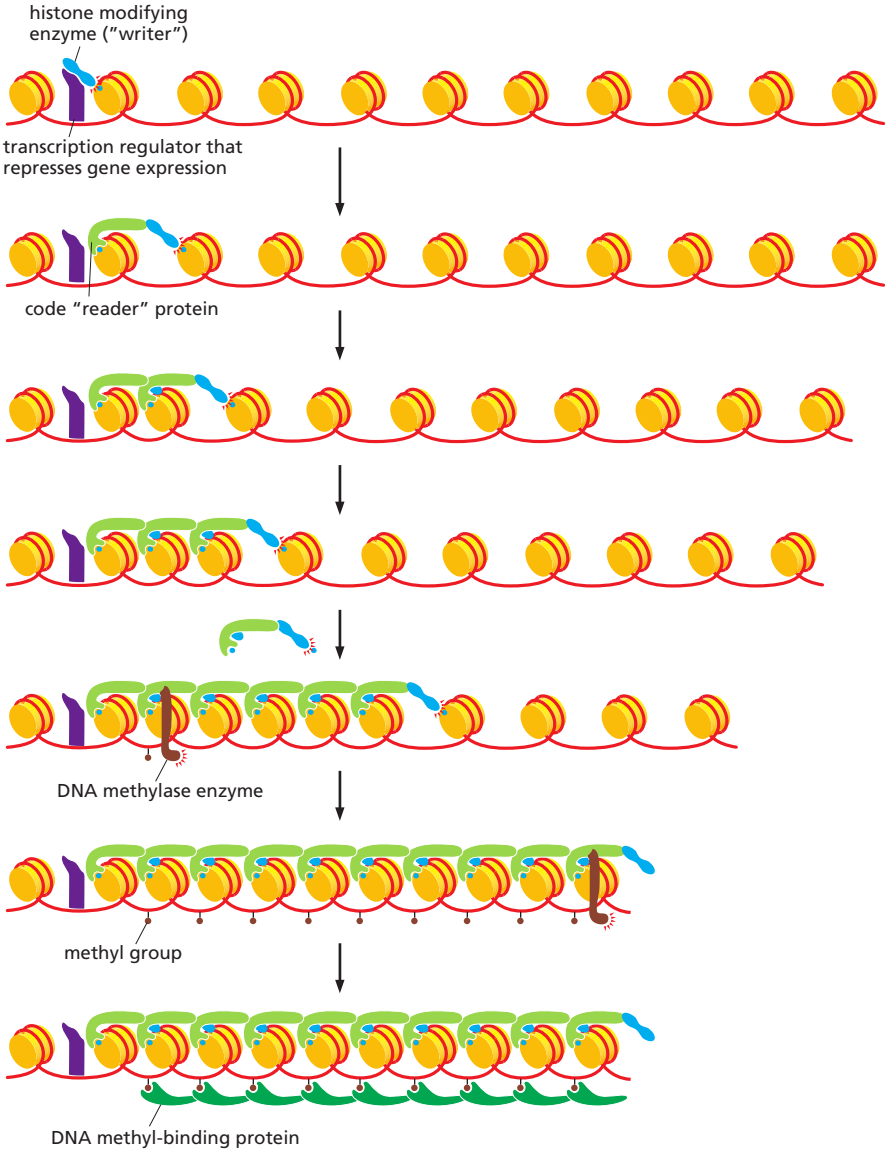


Figure 7–45 Multiple mechanisms contribute to stable gene repression. In this schematic example, histone reader and writer proteins (discussed in Chapter 4), under the direction of transcription regulators, establish a repressive form of chromatin. A *de novo* DNA methylase is attracted by the histone 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 (discussed in Chapter 4). At the same time, the mechanism shown in Figure 7–44 will cause both daughter chromosomes to inherit the same methylation pattern. In these cases where DNA methylation stimulates the activity of the histone writer, the two inheritance mechanisms will be mutually reinforcing. 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 4–44).

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 times their average density in selected regions, called **CG islands**, which average 1000 nucleotide pairs in length. The human genome contains roughly 20,000 CG islands and they usually include promoters of genes. For example, 60% of human protein-coding genes have promoters embedded in CG islands and these include virtually all 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 nearly all cells (**Figure 7–46**). Over evolutionary timescales, the CG islands were spared the accelerated mutation rate of bulk CG sequences because they remained unmethylated in the germ line (**Figure 7–47**).

CG islands also remain unmethylated in most somatic tissues whether or not the associated gene is expressed. The unmethylated state is maintained by sequence-specific DNA-binding proteins, many of whose *cis*-regulatory sequences contain a CG. By binding to these sequences, which are spread across CG islands, they protect the DNA from methyl transferases. These proteins also recruit *DNA demethylases*, which convert 5-methyl C to hydroxy-methyl C, which

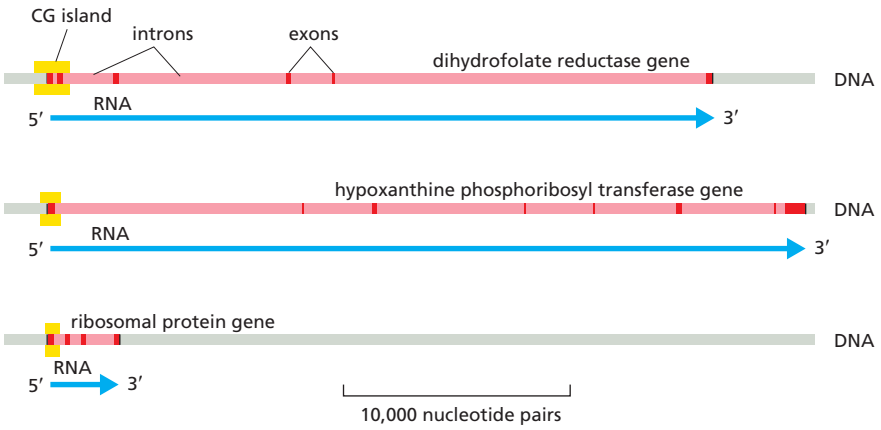


Figure 7–46 The CG islands surrounding the promoter in three mammalian housekeeping genes. The *yellow boxes* show the extent of each island. As for most genes in mammals, the exons *(dark red)* are very short relative to the introns *(light red)*. (Adapted from A.P. Bird, *Trends Genet.* 3:342–347, 1987. With permission from Elsevier.)

is later replaced by C either through DNA repair (see Figure 5–41A) or, passively, through multiple rounds of DNA replication. Unmethylated CG islands have several properties that make them particularly suitable for promoters. For example, some of the same proteins that bind to CG islands and protect them from methylation recruit histone modifying enzymes that make the islands particularly "promoter friendly." As a result, RNA polymerase is often found bound to promoters within CG islands, even when the associated gene is not being actively transcribed. At unmethylated CG islands, the balance between polymerase and nucleosome assembly is thus tipped toward the former. Additional steps are needed to "push" the bound polymerase into transcribing the adjacent gene, and these are directed by transcription regulators that bind to *cis*-regulatory sequences of DNA (often well upstream from the CG islands). These regulators serve to release the polymerase with the appropriate elongation factors (see Figure 7–21C and D).

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: when the paternally inherited gene copy is active, the maternally inherited gene copy is silent, or vice versa. This phenomenon is called **genomic imprinting**.

Roughly 300 genes are imprinted in humans. Because only one copy of an imprinted gene is expressed, imprinting can "unmask" mutations that would normally be covered by the other, functional copy. For example, Angelman syndrome, a disorder of the nervous system in humans that causes reduced mental ability and severe speech impairment, results from a gene deletion on one chromosomal homolog and the silencing, by imprinting, of the intact gene on the other homolog.

The *insulin-like growth factor-2* (*Igf2*) gene in the mouse provides a well-studied example of imprinting. 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.

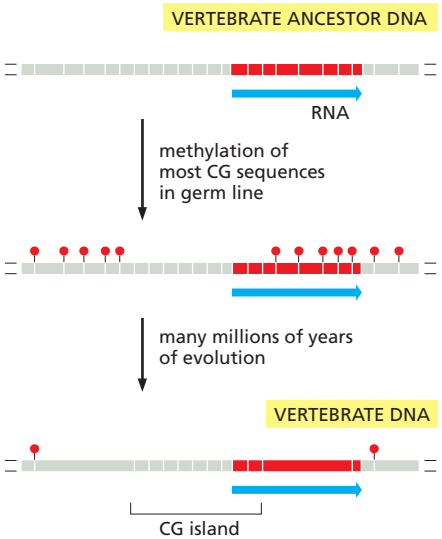


Figure 7–47 A mechanism to explain both the marked overall deficiency of CG sequences and their clustering into CG islands in vertebrate genomes. *White lines* mark the location of CG dinucleotides in the DNA sequences, while *red circles* indicate the presence of a methyl group on the CG dinucleotide. 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.

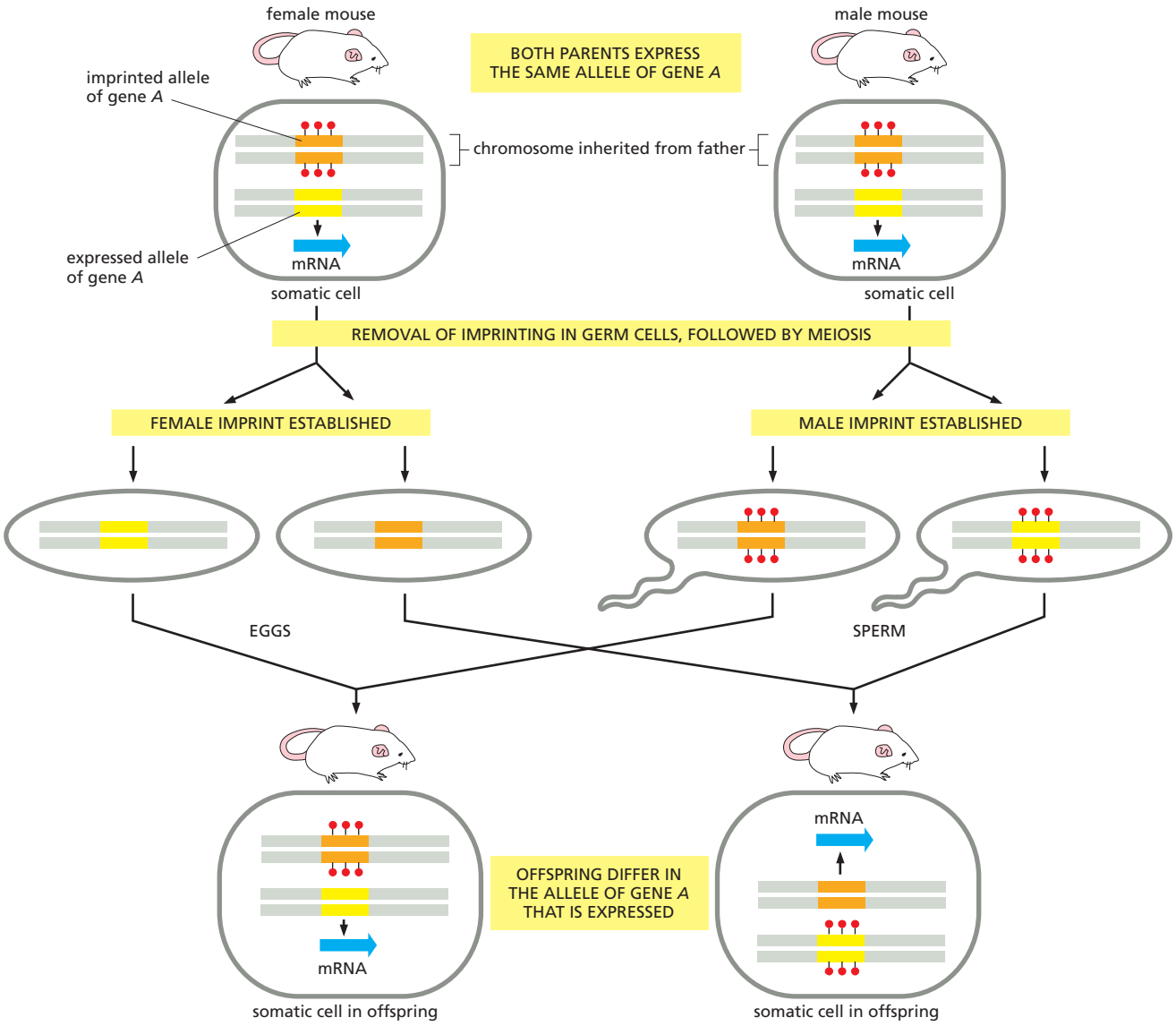
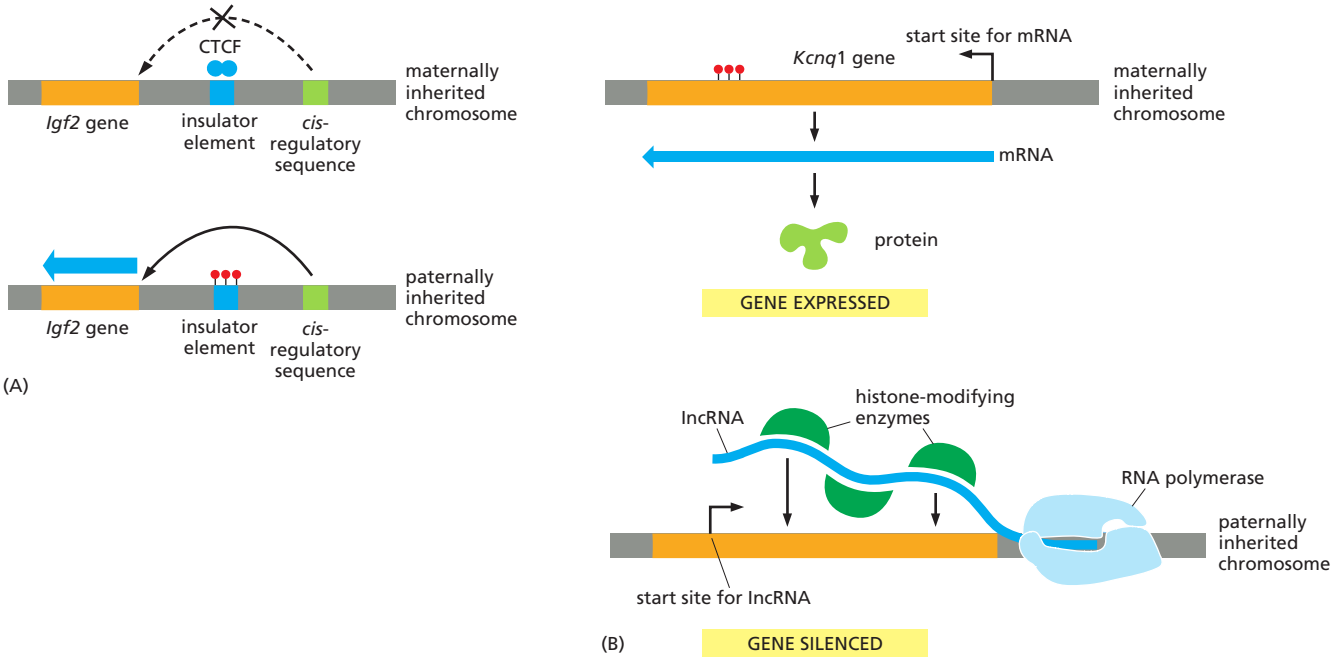


Figure 7–48 Imprinting in the mouse. The *top* portion of the figure shows a pair of homologous chromosomes in the somatic cells of two adult mice, one male and one female. In this example, both mice have inherited the top homolog from their father and the bottom homolog from their mother, and the paternal 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 gene *A* 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.

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 7–48). Because imprinted genes are somehow protected from the wave of demethylation that takes place shortly after fertilization (see pp. 404–405), this mark enables somatic cells to "remember" the 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, it can activate expression of a gene. In the case of *Igf2*, for example, methylation of an insulator element on the paternally derived chromosome blocks its function and allows distant *cis*-regulatory sequences 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 7–49**A).

Other cases of imprinting involve *long noncoding RNAs*, which are defined as RNA molecules over 200 nucleotides in length that do not code for proteins. We discuss lncRNAs broadly at the end of this chapter; here, we focus on the role of a specific lncRNA in imprinting. In the case of the *Kcnq1* gene, which codes for a voltage-gated calcium channel needed for proper heart function, the lncRNA is made from the paternal allele (which is unmethylated) but it is not released by the RNA polymerase, remaining instead at its site of synthesis on the DNA template. This RNA in turn recruits histone-modifying and DNA-methylating enzymes that direct the formation of repressive chromatin, which silences the protein-coding gene associated on the paternally derived chromosome (Figure 7–49B). The maternally derived gene, on the other hand, is immune to these effects because the specific methylation present from imprinting blocks the synthesis of the lncRNA but allows transcription of the protein-coding gene. Like *Igf2*, the specificity of *Kcnq1* imprinting arises from the inherited methylation patterns; the difference lies in the way these patterns bring about differential expression of the imprinted gene.

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.

Figure 7–49 Mechanisms of imprinting. (A) On chromosomes inherited from the female, a protein called CTCF binds to an insulator (see Figure 7–24), blocking communication between *cis*-regulatory sequences *(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 *(red circles)*; this inactivates the insulator by blocking the binding of the CTCF protein, and allows the *cis*-regulatory sequences to activate transcription of the *Igf2* gene. In other examples of imprinting, methylation simply blocks gene expression by interfering with the binding of proteins required for a gene's transcription. (B) Imprinting of the mouse *Kcnq1* gene. On the maternally derived chromosome, synthesis of the lncRNA is blocked by methylation of the DNA *(red circles)*, and the *Kcnq1* gene is expressed. On the paternally derived chromosome, the lncRNA is synthesized, remains in place, and by directing alterations in chromatin structure blocks expression of the *Kcnq1* gene. Although shown as directly binding to lncRNA, the histone-modifying enzymes are likely to be recruited indirectly, through additional proteins.

Chromosome-Wide Alterations in Chromatin Structure Can Be Inherited

We have seen that DNA methylation and certain types of chromatin structure can be heritable, preserving patterns of gene expression across 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 most 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. The correct ratio of X chromosome to *autosome* (non-sex chromosome) gene products is carefully controlled, and mutations that interfere with this dosage compensation are generally lethal.

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**. As a result of X-inactivation, two X chromosomes can coexist within the same nucleus, exposed to the same diffusible transcription regulators, yet differ entirely in their expression.

Early in the development of a female embryo, when it consists of a few hundred cells, one of the two X chromosomes in each cell becomes highly condensed into a type of heterochromatin. 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 hundred cells have already formed in the embryo, every female is a mosaic of clonal groups of cells in which either X_p or X_m is silenced (**Figure 7–50**). These clonal groups are

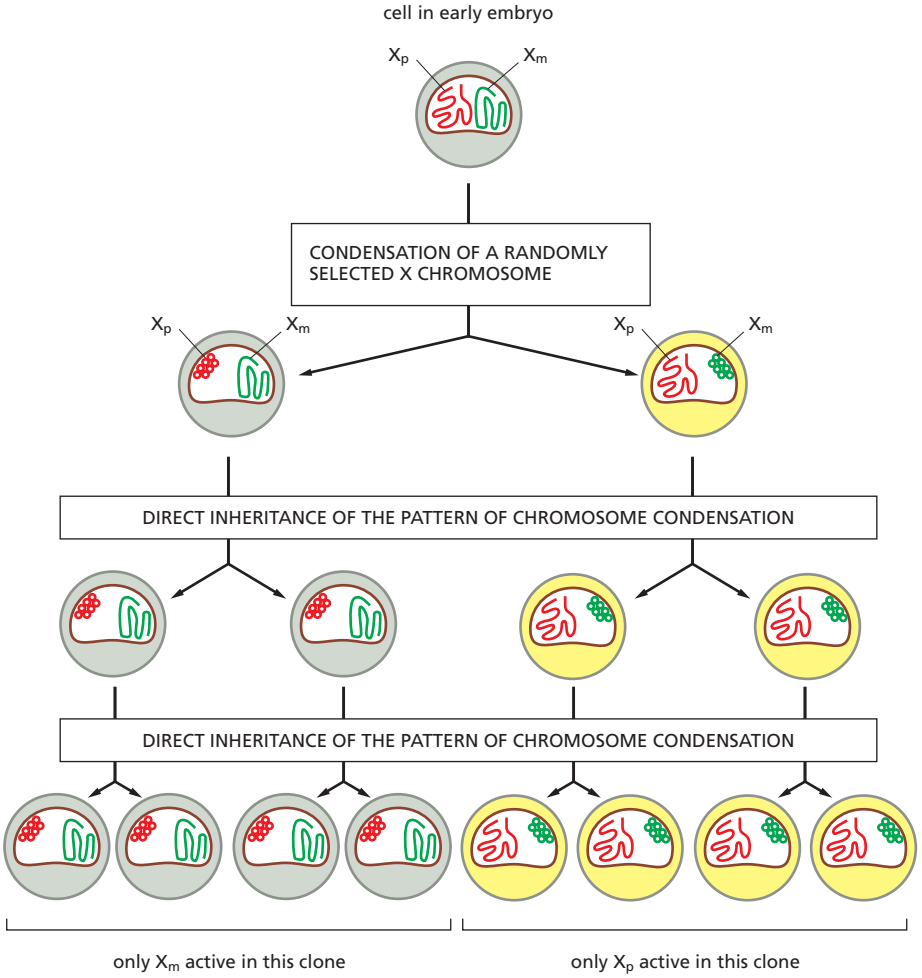


Figure 7–50 X-inactivation. The clonal inheritance in female mammals of a condensed, inactive X chromosome.

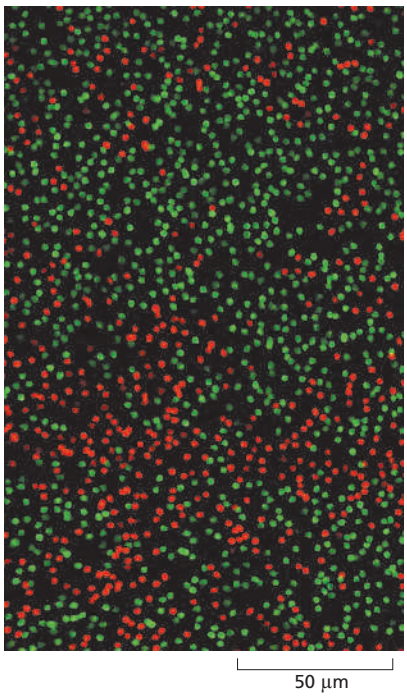


Figure 7–51 Photoreceptor cells in the retina of a female mouse showing patterns of X-chromosome inactivation. Using genetic engineering techniques (described in Chapter 8), the germ line of a mouse was modified so that one copy of the X chromosome (if active) makes a green fluorescent protein and the other a red fluorescent protein. Both proteins concentrate in the nucleus and, in the field of cells shown here, it is clear that only one of the two X chromosomes is active in each cell. (From H. Wu et al., *Neuron* 81:103–119, 2014. With permission from Elsevier.)

distributed in small clusters in the adult animal because sister cells tend to remain close together during later stages of development (**Figure 7–51**). For example, X-chromosome inactivation causes the orange and black "tortoiseshell" coat coloration of some female cats. In these cats, one X chromosome carries a gene that produces orange 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, male cats of this genetic stock are either solid orange 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 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 near the middle of the X chromosome, the **X-inactivation center** (**XIC**). Within the XIC is a transcribed 20,000-nucleotide lncRNA (called Xist), which is expressed solely from the inactive X chromosome. Xist RNA spreads from the XIC over the entire chromosome and directs gene silencing. Although we do not know exactly how this is accomplished, it likely involves recruitment of histone-modifying enzymes and other proteins to form a repressive form of chromatin analogous to that of Figure 7–45. Curiously, about 10% of the genes on the X chromosome (including *Xist* itself) escape this silencing and remain active.

The spread of Xist RNA along the X chromosome does not proceed linearly along the DNA. Rather, starting at its site of synthesis, it is first handed off across the base of the DNA loops that make up the chromosome; these shortcuts explain how Xist can spread rapidly, by a "hand-over-hand" mechanism, along the X chromosome once the inactivation process begins (**Figure 7–52**). It also helps to explain why the inactivation does not spread to the other, active X chromosome.

Imprinting and X-chromosome inactivation are examples of **monoallelic gene expression**, where in a diploid genome, only one of the two copies of a gene is expressed. In addition to the approximately 1000 genes on the X chromosome and the 300 or so genes that are imprinted, there are another 1000–2000 human genes that exhibit monoallelic expression. Like X-chromosome inactivation (but unlike imprinting), the choice of which copy of the gene is to be expressed and which is to be silenced often appears random. Yet once the choice is made, it can persist for many cell divisions. Because the choice is often made relatively late in development, cells of the same tissue in the same individual can express different copies of a given gene. In other words, somatic tissues are often mosaics, where different clones of cells have subtly different patterns of gene expression. The mechanisms responsible for this type of monoallelic expression are not known in detail, and its general purpose—if any—is poorly understood. Several different mechanisms may contribute to such epigenetic inheritance, as we explain next.

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. Perhaps the simplest way for a cell to remember

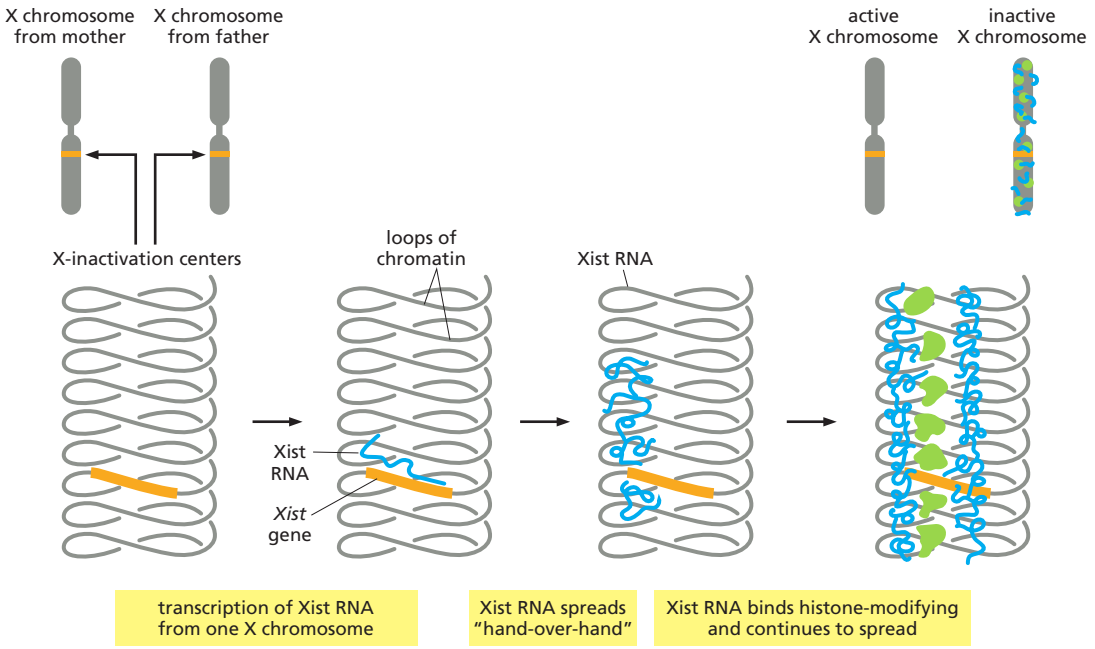


Figure 7–52 Mammalian X-chromosome inactivation. X-chromosome inactivation begins with the synthesis of Xist (X-inactivation specific transcript) RNA from the XIC (X-inactivation center) locus and moves outward to the chromosome ends. According to the model depicted here, the long (≈20,000 nucleotides) Xist RNA has many low-affinity binding sites for the structural components of chromosomes and spreads by releasing its hold on one portion of the chromosome while grasping another. The continued synthesis of Xist from the center of the chromosome drives it to the ends. As shown, Xist RNA does not move linearly along the chromosomal DNA, but, instead, moves first across the base of chromosome loops. It has been proposed that the portions of chromosomal DNA at the tips of long loops contain the 10% of genes that escape X-chromosome inactivation.

its identity is through a positive feedback loop in which a key transcription regulator activates, either directly or indirectly, the transcription of its own gene (see Figure 7–39). Interlocking positive feedback loops of the type shown in Figure 7–37 provide greater stability by buffering the circuit against fluctuations in the level of any one transcription regulator. Because transcription regulators are synthesized in the cytosol and diffuse throughout the nucleus, feedback loops based on this mechanism will affect both copies of a gene in a diploid cell. However, as discussed in this section, the expression pattern of a gene on one chromosome can differ from the copy of the same gene on the other chromosome (as in X-chromosome inactivation or in imprinting), and such differences can also be inherited through many cell divisions.

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**: a heritable alteration in a cell or organism's phenotype that does not result from changes in the nucleotide sequence of DNA (discussed in Chapter 4). (Unfortunately, the term epigenetic is sometimes also used to refer to all covalent modifications to histones and DNA, whether or not they are self-propagating; many of these modifications are erased each time a cell divides and do not generate cell memory.)

In **Figure 7–53**, we contrast two self-propagating epigenetic mechanisms that work in *cis,* affecting only one chromosomal copy with two self-propagating mechanisms that work in *trans,* affecting both chromosomal copies of a gene. Cells can 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.

We can get some idea of the prevalence of epigenetic changes by comparing identical twins. Their genomes have the same sequence of nucleotides, and, obviously, many features of identical twins—such as their appearance—are strongly determined by the genome sequences they inherit. When their gene expression, histone modification, and DNA methylation patterns are compared, however, 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 has been proposed that some of these differences are heritable from cell to cell and are the result of environmental factors. Although these studies are in early stages, the idea that environmental events can be permanently registered as epigenetic changes in our cells is a fascinating one that presents an important challenge to the next generation of biological scientists.

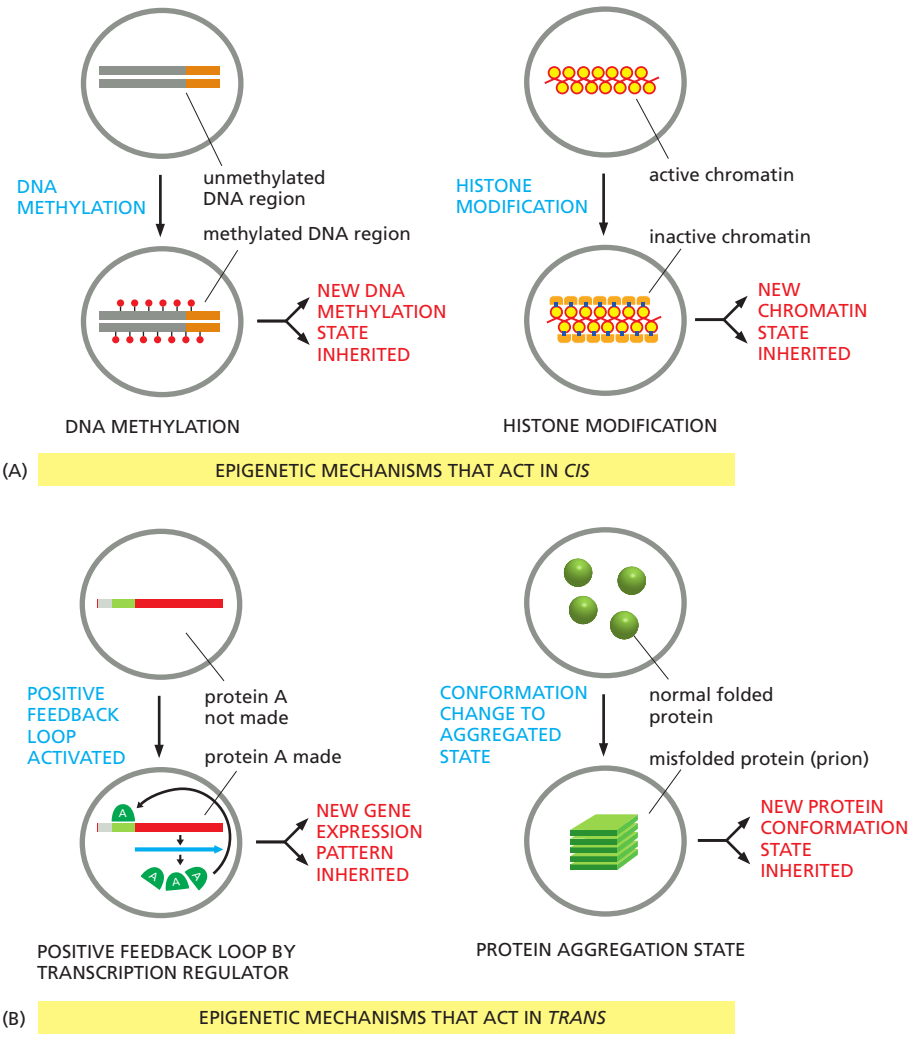


Figure 7–53 Four distinct mechanisms that can produce an epigenetic form of inheritance in an organism. (A) Epigenetic mechanisms that act in *cis*. As discussed in this chapter, a maintenance methylase can propagate specific patterns of cytosine methylation (see Figure 7–44). As discussed in Chapter 4, a histone modifying enzyme that replicates the same modification that attracts it to chromatin can result in the modification being self-propagating (see Figure 4–44). (B) Epigenetic mechanisms that act in *trans*. Positive feedback loops, formed by transcriptional regulators are found in all species and are probably the most common form of cell memory. As discussed in Chapter 3, some proteins can form self-propagating prions (Figure 3–33). If these proteins are involved in gene expression, they can transmit patterns of gene expression to daughter cells.

Summary

Eukaryotic cells can use inherited forms of DNA methylation and inherited states of chromatin condensation as additional mechanisms for generating cell memory of gene expression patterns. An especially dramatic case that involves chromatin condensation is the inactivation of an entire X chromosome in female mammals. DNA methylation underlies the phenomenon in mammals of genomic imprinting, 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 has 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 7–54).

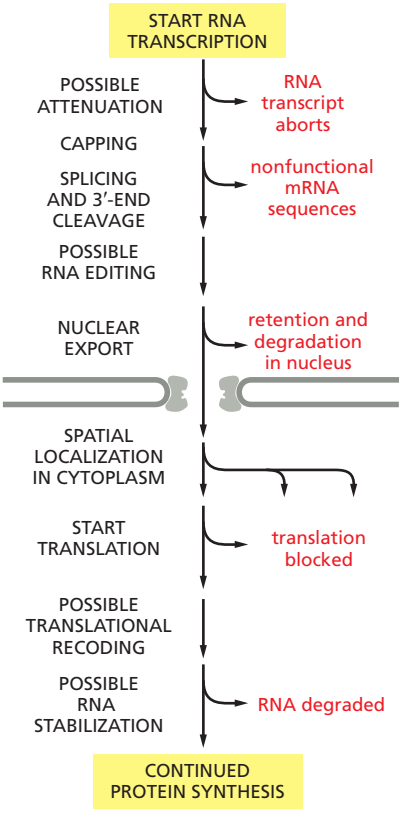
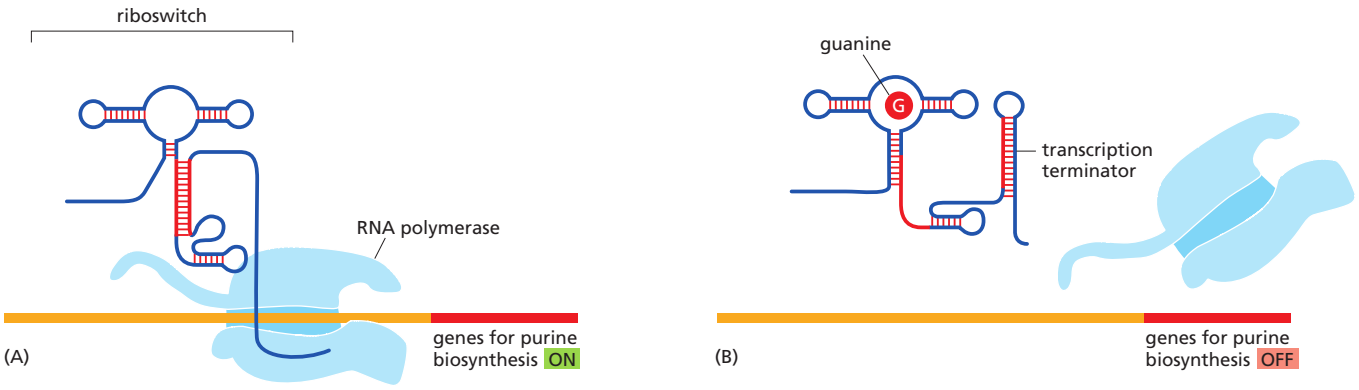


Figure 7–54 Post-transcriptional controls of gene expression. The final synthesis rate of a protein can, in principle, be controlled at any of the steps listed in capital letters. In addition, RNA splicing, RNA editing, and translation recoding can also alter the sequence of amino acids in a protein, 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

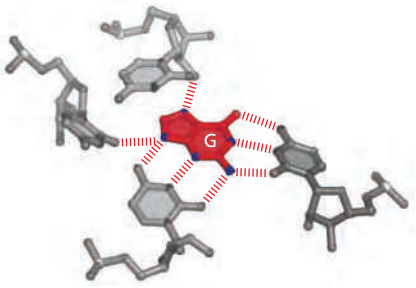
It has long been known that the expression of some genes is inhibited by premature termination of transcription, a phenomenon called *transcription attenuation*. In some of these cases, the nascent RNA chain adopts a structure that causes it to interact with the RNA polymerase in such a way as to abort its transcription. When the gene product is required, regulatory proteins bind to the nascent RNA chain and remove the attenuation, allowing the transcription of a complete RNA molecule.

A well-studied example of transcription attenuation occurs during the life cycle of HIV, the human immunodeficiency virus that is the causative agent of acquired immune deficiency syndrome, or AIDS. Once the HIV genome has been integrated into the host genome, the viral DNA is transcribed by the cell's RNA polymerase II (see Figure 5–62). However, this polymerase usually terminates transcription after synthesizing transcripts of several hundred nucleotides and therefore fails to 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 (see Figure 6–89). Once bound to this specific RNA structure (called TAR), Tat assembles several host-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. Thus, a normal cell mechanism has apparently been highjacked by HIV to permit transcription of its genome to be controlled by a single viral protein.

Riboswitches Probably Represent Ancient Forms of Gene Control

In Chapter 6, we discussed the idea that, before modern cells arose on Earth, RNA played the role of both DNA and proteins, both storing hereditary information and catalyzing chemical reactions (see pp. 362–366). The discovery of *riboswitches* shows that RNA can also form control devices. 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 7–55**).

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 7–55C). Moreover, the binding affinities observed are as tight as those typically observed between small molecules and proteins.



(C)

Figure 7–55 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*-adenosylmethionine, coenzyme B_{12}, flavin mononucleotide, adenine, lysine, and glycine. (Adapted from M. Mandal and R.R. Breaker, *Nat. Rev. Mol. Cell Biol.* 5:451–463, 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.)

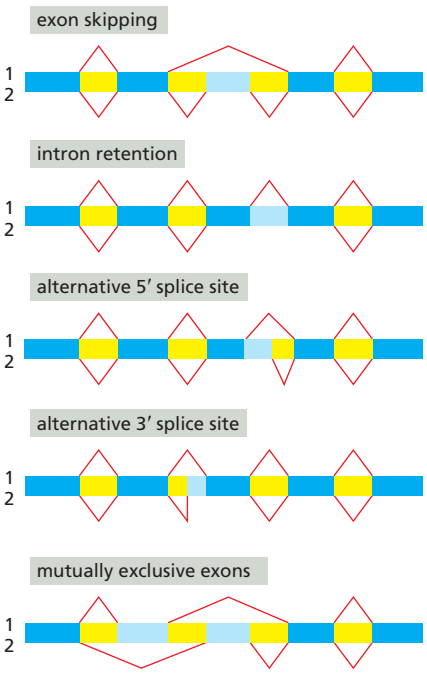


Figure 7–56 Five patterns of alternative RNA splicing. In each case, a single type of RNA transcript is spliced in two alternative ways to produce two distinct mRNAs (1 and 2). The *dark blue boxes* mark exon sequences that are retained in both mRNAs. The *light blue boxes* mark possible exon sequences 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 from H. Keren et al. *Nat. Rev. Genet*. 11:345–355, 2010. With permission from Macmillan Publishers Ltd.)

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 7–55, the riboswitch controls transcription elongation, but they can 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, a fact that supports the hypothesis of an early "RNA world."

Alternative RNA Splicing Can Produce Different Forms of a Protein from the Same Gene

As discussed in Chapter 6 (see Figure 6–26), RNA splicing shortens the transcripts of many eukaryotic 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** (**Figure 7–56**). A substantial proportion of animal genes (estimated at 90% 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 7–57**), 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 rare in single-celled budding yeasts

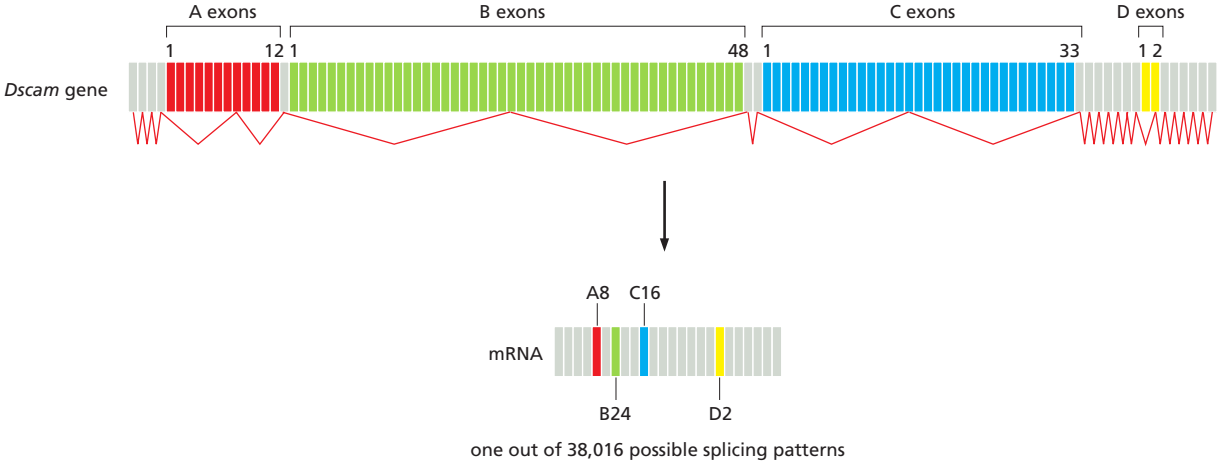


Figure 7–57 Alternative splicing of RNA transcripts of the *Drosophila Dscam* gene. DSCAM proteins have several different functions. In cells of the fly immune system, they mediate the phagocytosis of bacterial pathogens. In cells of the nervous system, DSCAM proteins are needed for proper wiring of neurons. 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)*. 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 24–48), but the amino acid sequence of the domains vary according to the splicing pattern. The diversity of DSCAM variants contributes to the plasticity of the immune system as well as the formation of complex neural circuits; we take up the specific role of the DSCAM variants in more detail when we describe the development of the nervous system in Chapter 21. (Adapted from D.L. Black, *Cell* 103:367–370, 2000. With permission from Elsevier.)

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.

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 clearly between two or more alternative pairings of 5′ and 3′ splice sites, so that different choices are made by chance on different individual 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. In the simplest examples, regulated splicing is used to switch from the production of a nonfunctional protein to the production of a functional one (or the other way around). 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 5–61). 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 enabling switching from the production of a functional protein to the production of a nonfunctional one (or vice versa), 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 6–26). 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 7–58**).

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. 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 effects on splicing of regulatory proteins.

The Definition of a Gene Has Been Modified Since the Discovery of Alternative RNA Splicing

The discovery that eukaryotic 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*.

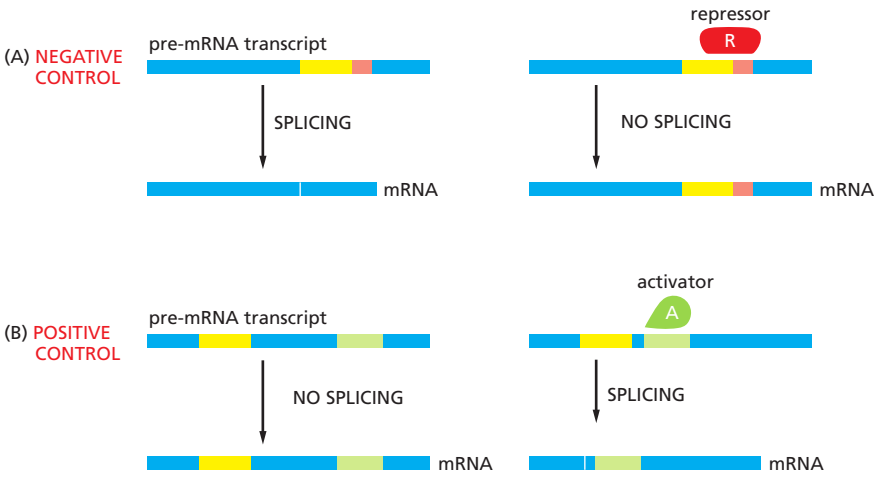


Figure 7–58 Negative and positive control of alternative RNA splicing. (A) In negative control, a repressor protein binds to a specific sequence in the pre-mRNA transcript and blocks access of the splicing machinery to a splice junction. This often results in the use of a secondary splice site, thereby producing an altered pattern of splicing (see Figure 7–56). (B) In positive control, the splicing machinery is unable to remove a particular intron sequence efficiently without assistance from an activator protein. Because RNA is flexible, the nucleotide sequences that bind these activators can be located many nucleotide pairs from the splice junctions they control, and they are often called *splicing enhancers*, by analogy with the transcriptional enhancers mentioned earlier in this chapter.

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 eukaryotic 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 eukaryotic 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 a DNA sequence that is transcribed as a single unit and encodes one set of closely related polypeptide chains (protein isoforms) as a single protein-coding gene. This definition also accommodates those DNA sequences that encode protein variants produced by post-transcriptional processes other than RNA splicing, such as transcript cleavage and RNA editing (discussed below).

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 eukaryotic mRNA molecule is not formed by the termination of RNA synthesis by the RNA polymerase, as it is in bacteria. Instead, it results from an RNA cleavage reaction that is catalyzed by additional proteins while the transcript is elongating (see Figure 6–34). A cell can control the site of this cleavage so as to change the C-terminus of the resultant protein. In the simplest cases, one protein variant is simply a truncated version of the other; in many other cases, however, the alternative cleavage and polyadenylation sites lie within intron sequences and the pattern of splicing is thereby altered. This process can produce two closely related proteins differing only in the amino acid sequences at their C-terminal ends. Close analysis of RNAs produced from the human genome in a variety of cell types (see Figure 7–3) indicate that as many as 50% of human protein-coding genes produce mRNA species that differ at their site of polyadenylation.

A well-studied example of regulated polyadenylation is the switch from the synthesis of membrane-bound to secreted antibody molecules that occurs during the development of B lymphocytes (see Figure 24–22). 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 is generated through a change in the site of RNA cleavage and polyadenylation, as shown in **Figure 7–59**.

The change is caused by an increase in the concentration of a subunit of a protein (CstF) that promotes RNA cleavage (see Figure 6–34). The first cleavage/poly-A addition site that a transcribing RNA polymerase encounters is suboptimal and is usually skipped in unstimulated B lymphocytes, leading to production

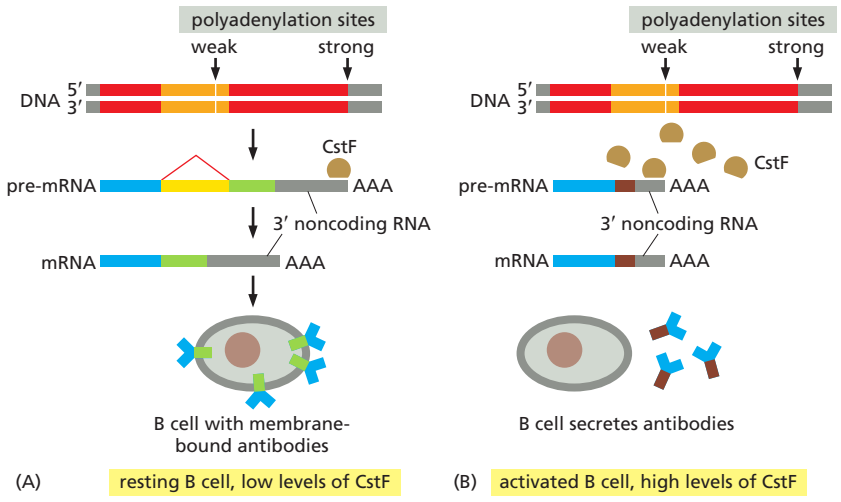


Figure 7–59 Regulation of the site of RNA cleavage and poly-A addition determines whether an antibody molecule is secreted or remains membrane-bound. In unstimulated B lymphocytes *(left)*, a long RNA transcript is produced, and the intron sequence *(yellow)* near its 3′ end is removed by RNA splicing to provide an mRNA molecule that codes for a membrane-bound antibody molecule. Only a portion of the antibody gene is shown in the figure; the actual gene and its mRNA would extend further to the left of the diagram. After antigen stimulation *(right)*, the RNA transcript is cleaved and polyadenylated upstream from the intron's 3′ splice site. As a result, some of the intron sequence remains as a coding sequence in the short transcript and specifies the hydrophilic C-terminal portion of the secreted antibody molecule *(brown)*. (Adapted from D. Di Giammartino et al., *Mol. Cell* 43:853–866, 2011. With permission from Elsevier.)

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 has 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. We saw in Chapter 6 that tRNA and rRNA molecules are chemically modified after they are synthesized: here we focus on changes to mRNAs.

In animals, two principal types of mRNA editing occur: the deamination of adenine to produce inosine (A-to-I editing) and, less frequently, the deamination of cytosine to produce uracil (C-to-U editing), as shown in Figure 5–43. 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 by creating a premature stop codon. Edits that occur outside coding sequences can affect the pattern of pre-mRNA splicing, the transport of mRNA from the nucleus to the cytosol, the efficiency with which the RNA is translated, or the base-pairing between microRNAs (miRNAs) and their mRNA targets, a form of regulation that will be discussed later in the chapter.

The process of A-to-I editing is particularly prevalent in humans, where it occurs in approximately 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 an intron (**Figure 7–60**). The structure of the double-stranded RNA specifies whether the mRNA is to be edited, and if so, 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. Mutant mice that cannot make this edit are prone to epileptic seizures and die during or shortly after weaning, 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

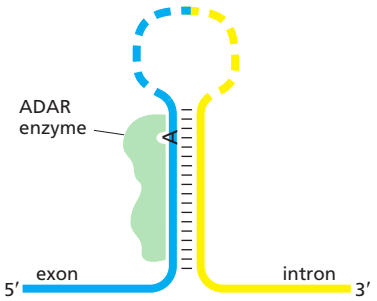


Figure 7–60 Mechanism of A-to-I RNA editing in mammals. Typically, a sequence complementary to the position of the edit is present in an intron, and the resulting double-stranded RNA structure attracts an A-to-I editing enzyme (ADAR). In the case illustrated, the edit is made in an exon; in most cases, however, this occurs in noncoding portions of the mRNA. Editing by ADAR takes place in the nucleus, before the pre-mRNA has been fully processed. Mice and humans have two ADAR genes: *ADR1* is expressed in many tissues and is required in the liver for proper red blood cell development; *ADR2* is expressed only in the brain, where it is required for proper brain development.

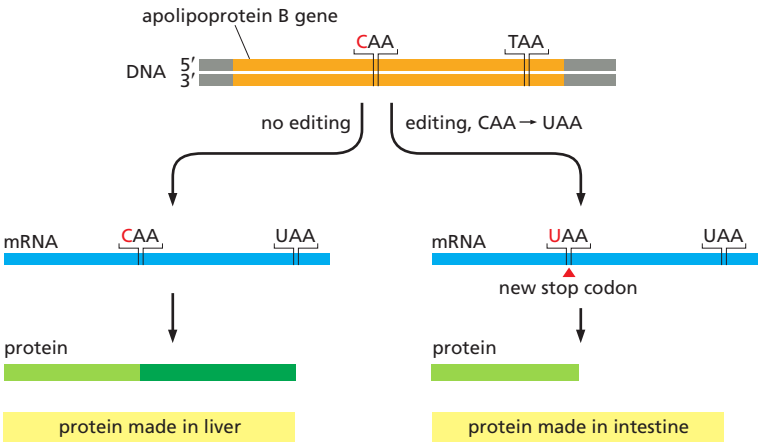


Figure 7–61 C-to-U RNA editing produces a truncated form of apolipoprotein B.

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 role in lipid metabolism that is specific to the organ that produces it (**Figure 7–61**).

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 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, 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, RNA editing presumably helps to hold many viruses in check.

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 for 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 5–62). Once inserted, the viral DNA can be 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 7–62**). 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. 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.

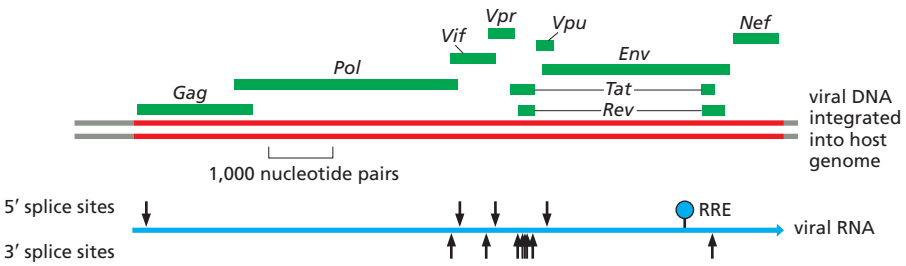


Figure 7–62 The compact genome of HIV, the human AIDS virus. The positions of the nine HIV genes are shown in *green.* The *red double line* indicates a DNA copy of the viral genome that has become integrated into the host DNA *(gray).* Note that the coding regions of many genes overlap, and that those 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. The *Env* gene codes for the envelope proteins (see Figure 5–62). Tat, Rev, Vif, Vpr, Vpu, and Nef are small proteins with a variety of functions. For example, Rev regulates nuclear export (see Figure 7–63) and Tat regulates the elongation of transcription across the integrated viral genome (see p. 414).

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 (Crm1), 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.

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 the 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 7–63**). Regulation by Rev and by Tat, the HIV protein that counteracts premature transcription termination (see p. 414), allows the 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.

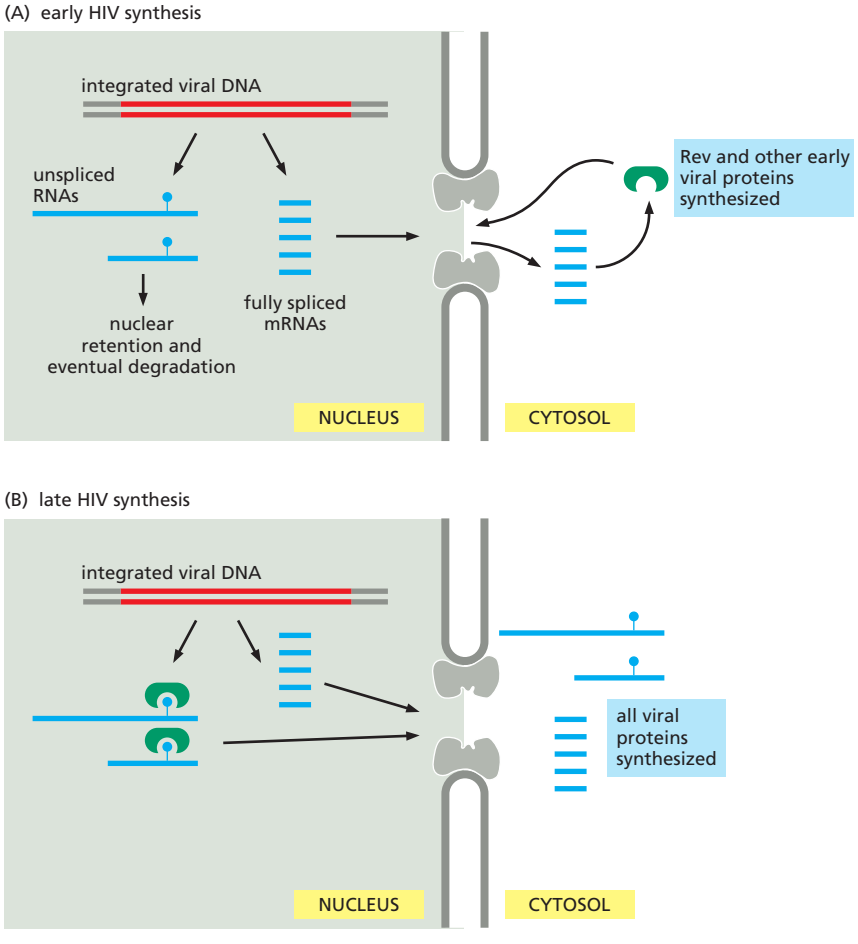


Figure 7–63 Regulation of nuclear export by the HIV Rev protein. (A) Early in HIV infection, only the fully spliced RNAs (which contain the coding sequences for Rev, Tat, and Nef) are exported from the nucleus and translated. (B) Once sufficient Rev protein has accumulated and been transported into the nucleus, 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.

If, after its initial entry into a host cell, conditions become unfavorable for viral transcription and replication, Rev and Tat are made at levels too low to promote transcription and export of unspliced RNA. This situation stalls the viral growth cycle until conditions improve, whereupon Rev and Tat levels increase, and the virus enters the replication cycle.

Some mRNAs Are Localized to Specific Regions of the Cytosol

Once a newly made eukaryotic 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 6–8). Once the first round of translation "passes" the nonsense-mediated decay test (see Figure 6–76), 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 N-terminus will direct it to the endoplasmic reticulum (ER). In this case, as discussed in Chapter 12, 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. In other cases, free 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.

Many 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. 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. This strategy also provides the cell with other 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 translational control, also allows the cell to regulate gene expression independently in different regions. This feature is particularly important in large, highly polarized cells such as neurons, where it plays a central role in synaptic function.

Several mechanisms for mRNA localization have been discovered (**Figure 7–64**), all of which require specific signals in the mRNA itself. These signals are usually concentrated in the 3′ *untranslated region* (*UTR*), the region of RNA that

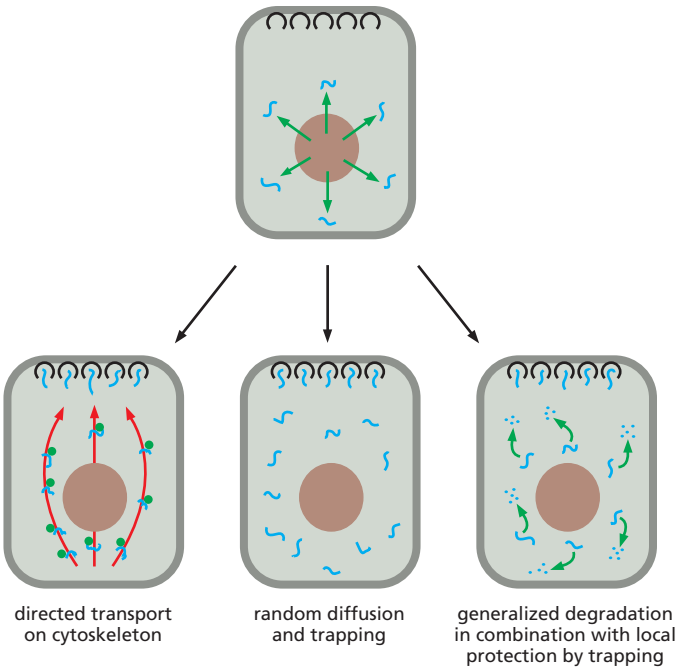


Figure 7–64 Mechanisms for the localization of mRNAs. The mRNA to be localized leaves the nucleus through nuclear pores *(top).* Some localized mRNAs *(left diagram)* travel to their destination by associating with cytoskeletal motors, which use the energy of ATP hydrolysis to move the mRNAs unidirectionally along filaments in the cytoskeleton *(red)* (see Chapter 16). At their destination, the mRNAs are held in place by anchor proteins *(black).* Other mRNAs randomly diffuse through the cytosol and are simply trapped by anchor proteins and at their sites of localization *(center diagram).* Some 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 mechanism requires specific signals on the mRNA, which are typically located in the 3′ UTR. Additional components can 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.)

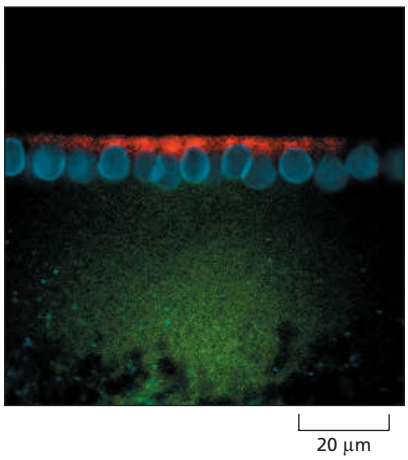


Figure 7–65 An experiment demonstrating the importance of the 3′ UTR in localizing mRNAs to specific regions of the cytoplasm. For this experiment, two different fluorescently labeled RNAs were prepared by transcribing DNA *in vitro* in the presence of fluorescently labeled derivatives of UTP. One RNA (labeled with a *red* fluorochrome) contains the coding region for the *Drosophila* Hairy protein and includes the adjacent 3′ UTR (see Figure 6–21). 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 7–26). 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 transcriptional regulators that specify positional information in the developing *Drosophila* embryo (discussed in Chapter 21), and the localization of its mRNA (shown in this experiment to depend on its 3′ UTR) is critical for proper fly development. (Courtesy of Simon Bullock and David Ish-Horowicz.)

extends from the stop codon that terminates protein synthesis to the start of the poly-A tail (**Figure 7–65**). The 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 transcription regulator 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 (see Figure 7–26). 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 eukaryotic 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 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 eukaryotes 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 nucleotides, the *Shine–Dalgarno sequence*, is always found a few nucleotides upstream of the initiating AUG codon. In bacteria, translational control mechanisms are carried out by proteins or by RNA molecules, and they generally involve either exposing or blocking the Shine–Dalgarno sequence (**Figure 7–66**).

Eukaryotic mRNAs do not contain such a 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. In eukaryotes, translational repressors can 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 6–70). A particularly important type of translational control in eukaryotes relies on small RNAs (termed *microRNA*s or *miRNA*s) that bind to mRNAs and reduce protein output, as described later in this chapter.

The Phosphorylation of an Initiation Factor Regulates Protein Synthesis Globally

Eukaryotic cells decrease their overall rate of protein synthesis in response to a variety of 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.

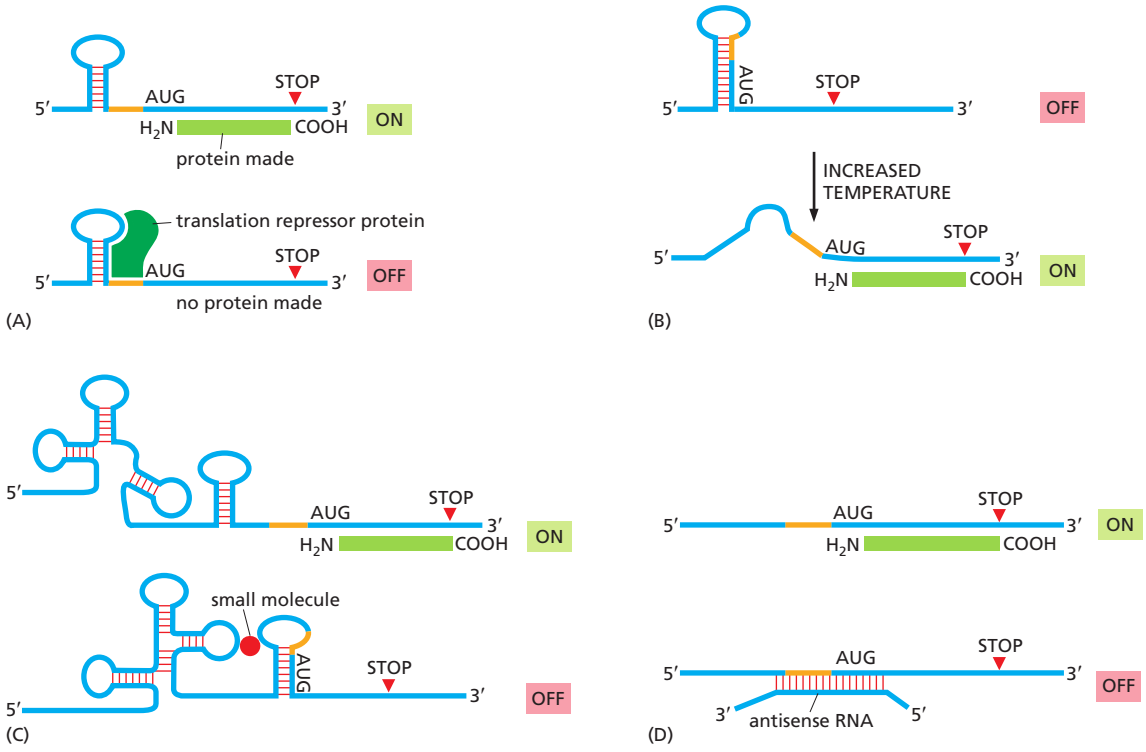


Figure 7–66 Mechanisms of translational control. Although these examples are from bacteria, many of the same principles operate in eukaryotes. (A) Sequence-specific RNA-binding proteins repress translation of specific mRNAs by blocking access of the ribosome to the Shine–Dalgarno sequence *(orange)*. For example, some ribosomal proteins repress translation of their own RNA. This mechanism allows the cell to maintain correctly balanced quantities of the various components needed to form ribosomes. (B) An RNA "thermosensor" permits efficient translation initiation only at elevated temperatures at 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 major rearrangement of the RNA forming a different set of stem-loop structures. In the bound structure, the Shine–Dalgarno sequence *(orange)* is sequestered and translation initiation is thereby blocked. In many bacteria, *S*-adenosylmethionine 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.

Because eIF2 binds very tightly to GDP, a guanine nucleotide exchange factor (see p. 157), designated eIF2B, is required to cause GDP release so that a new GTP molecule can bind and eIF2 can be reused (**Figure 7–67**A). The 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 7–67B).

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.

Initiation at AUG Codons Upstream of the Translation Start Can Regulate Eukaryotic Translation Initiation

We saw in Chapter 6 that eukaryotic 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 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 N-termini, from the same mRNA. A particularly important use of this mechanism is the production of the same protein with and without a signal sequence attached at its N-terminus. This allows the protein to be directed to two different locations in the cell (for example, to both mitochondria and the cytosol). Cells 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 eukaryotes uses one or more short *open reading frames*—short stretches of DNA that begin with a start codon (ATG) and end with a stop codon, with no stop codons in between—that lie between the 5′ end of the mRNA and the beginning of the gene. 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

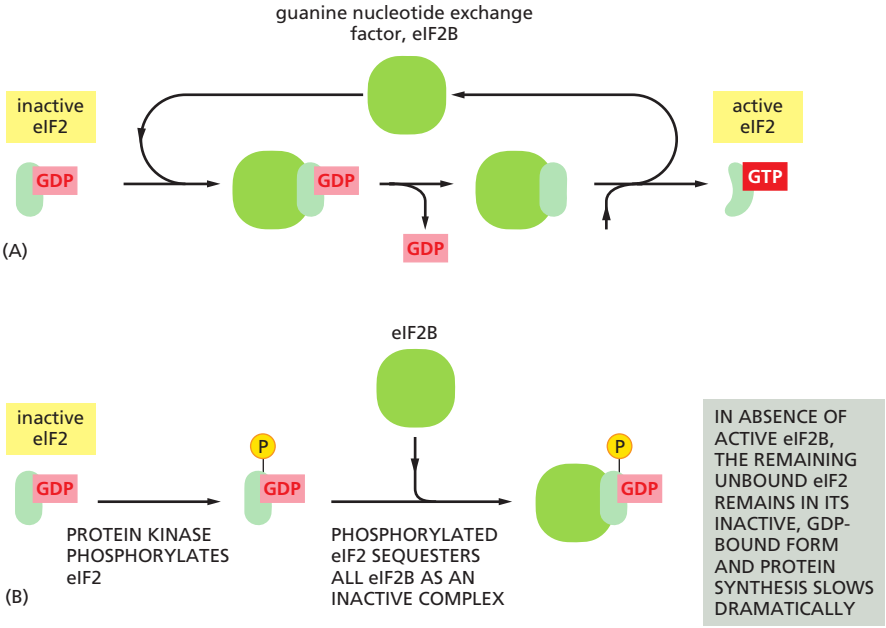


Figure 7–67 The eIF2 cycle. (A) The recycling of used eIF2 by a guanine nucleotide exchange factor (eIF2B). (B) eIF2 phosphorylation controls protein synthesis rates by tying up eIF2B.

scanning ribosome initiation complex and causing the ribosome to translate the uORF and dissociate from the mRNA before it reaches the bona fide protein-coding sequence.

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 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 missing nutrients. The details of this mechanism have been worked out for a specific yeast mRNA that encodes a protein called Gcn4, a transcription regulator that activates many genes that encode proteins that are important for amino acid synthesis.

The *Gcn4* mRNA contains several short uORFs, and when amino acids are abundant, ribosomes translate the uORFs and generally dissociate before they reach the *Gcn4* coding region. A global decrease in eIF2 activity brought about by amino acid starvation makes it more likely that a scanning small ribosomal subunit will move across the uORFs (without translating them) before it acquires a molecule of eIF2 (see Figure 6–70). Such a ribosomal subunit is then free to initiate translation on the actual *Gcn4* sequences. The increased level of this transcription regulator increases production of the amino acid biosynthetic enzymes.

Internal Ribosome Entry Sites Provide Opportunities for Translational Control

Although approximately 90% of eukaryotic 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)**. In some cases, two distinct protein-coding sequences are carried in tandem on the same eukaryotic 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 7–68**). 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.

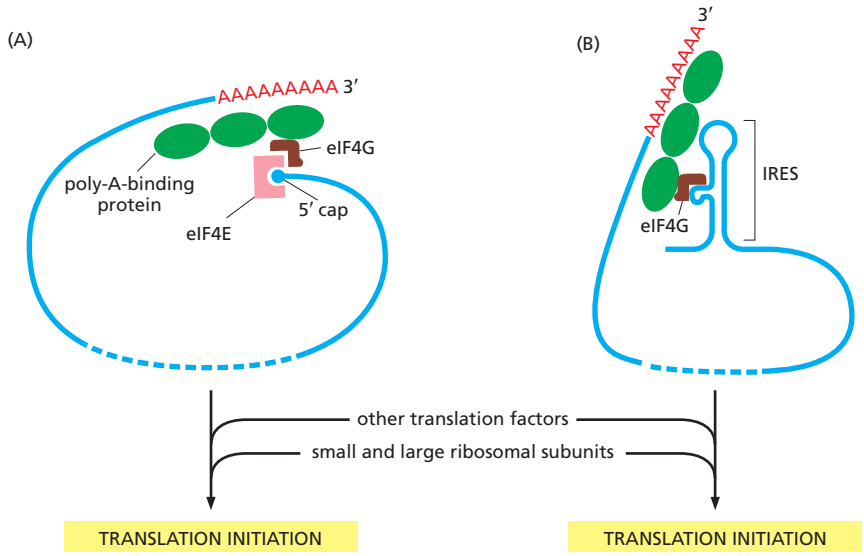


Figure 7–68 Two mechanisms of translation initiation. (A) The normal, cap-dependent mechanism requires a set of initiation factors whose assembly on the mRNA is stimulated by the presence of a 5′ cap and a poly-A tail (see also Figure 6–70). (B) The IRES-dependent mechanism, seen mainly in viruses, requires 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.)

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 present on the viral mRNAs. (The truncated eIF4G remains competent to initiate translation at these internal sites.)

The many ways in which viruses manipulate their host's protein-synthesis machinery for their own advantage continue to surprise cell biologists. Studying this "arms race" between humans and pathogens has led to many fundamental insights into the workings of the cell, and we revisit this topic in more detail in Chapter 23.

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 eukaryotic 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 transcription regulators, whose production rates need to change rapidly in cells, have especially short half-lives.

We saw in Chapter 6 that the cell has several mechanisms that rapidly destroy incorrectly processed RNAs; here, we consider the fate of the typical "normal" eukaryotic mRNA. Two general mechanisms exist for eventually destroying each mRNA that is made by the cell. 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 cytosol. 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 7–69**).

Nearly all mRNAs are subject to both 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 7–70**).

Although poly-A shortening controls the half-life of most eukaryotic mRNAs, some mRNAs can be degraded by a specialized mechanism that bypasses this step altogether. In these cases, specific nucleases cleave the mRNA internally,

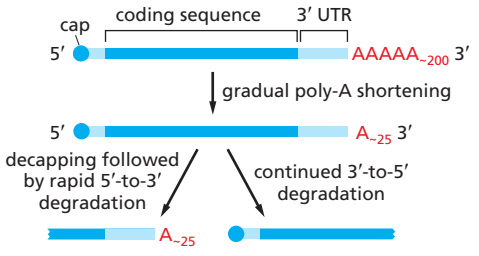


Figure 7–69 Two mechanisms of eukaryotic mRNA decay. A critical threshold of poly-A tail length induces rapid 3′-to-5′ degradation, which may be triggered by the loss of the poly-A-binding proteins. As shown in Figure 7–70, a deadenylase associates with both the 3′ poly-A tail and the 5′ cap, and this connection may be involved in signaling 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.)

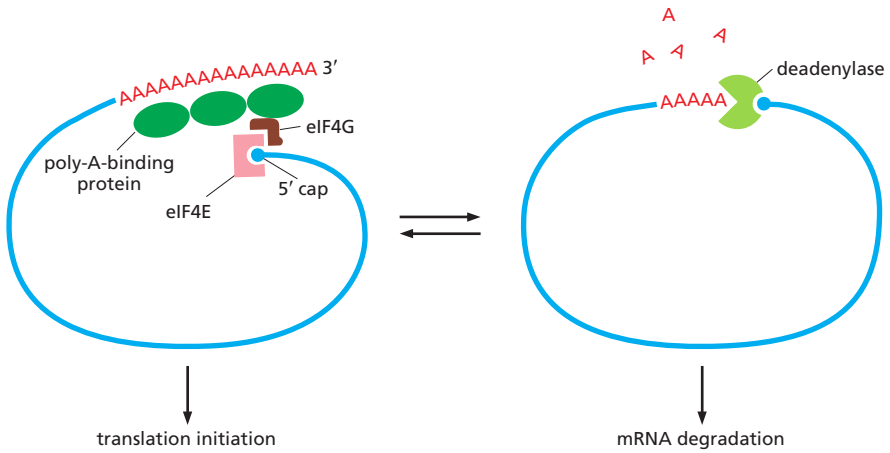


Figure 7–70 The competition between mRNA translation and mRNA decay. The same two features of an mRNA molecule—its 5′ cap and the 3′ poly-A tail—are used in both translation initiation and deadenylation-dependent mRNA decay (see Figure 7–69). The deadenylase that shortens the poly-A tail in the 3′-to-5′ direction associates with the 5′ cap. As described in Chapter 6 (see Figure 6–70), 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.)

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 or exposing 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. Aconitase can bind to the 3′ UTR of the transferrin receptor mRNA and increase receptor production by 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 7–71**).

Regulation of mRNA Stability Involves P-bodies and Stress Granules

We saw in Chapters 3 and 6 that large aggregates of proteins and nucleic acids that work together are often held in proximity by loose, low-affinity connections (see Figure 3–36). In this way, they function as "organelles" even though they are not surrounded by membranes. Many of the events discussed in the previous

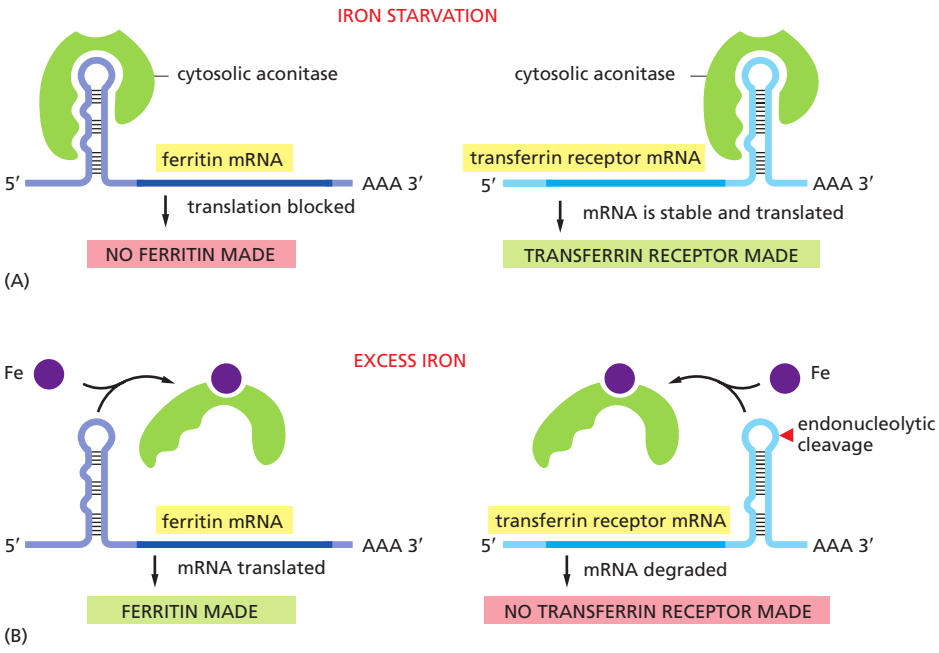


Figure 7–71 Two post-translational controls mediated by iron. (A) During iron starvation, the binding of aconitase to the 5′ UTR of the ferritin mRNA blocks translation initiation; its binding to the 3′ UTR of the transferrin receptor mRNA blocks an endonuclease cleavage site and thereby stabilizes the 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-loop structure in the mRNAs encoding ferritin and the 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.)

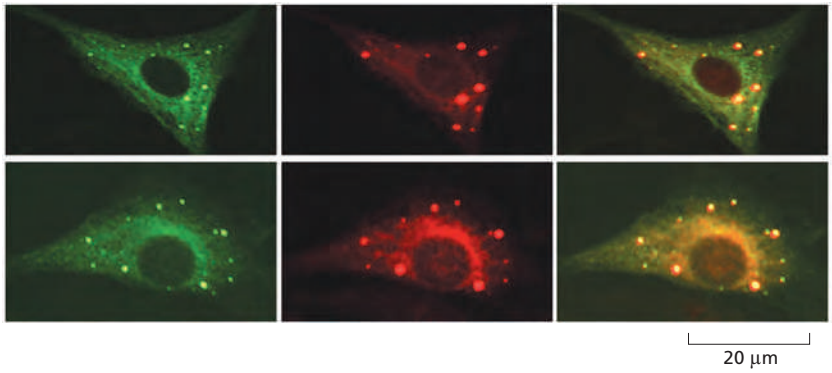


Figure 7–72 Visualization of P-bodies. Human cells were stained with antibodies to a component of the mRNA decapping enzyme Dcp1a *(left panels)* and to the Argonaute protein *(middle panels).* As described later in this chapter, Argonaute is a key component of RNA interference pathways. The merged image *(right panels)* shows that the two proteins co-localize to P-bodies in the cytoplasm. (Adapted from J. Liu et al., *Nat. Cell Biol.* 7:719–723, 2005. With permission from Macmillan Publishers Ltd.)

section—including decapping and RNA degradation—take place in aggregates known as *Processing-* or *P-bodies*, which are present in the cytosol (**Figure 7–72**).

Although many mRNAs are eventually degraded in P-bodies, some remain intact and are later returned to the pool of translating mRNAs. To be "rescued" in this way, mRNAs move from P-bodies to another type of aggregate known as a *stress granule*, which contains translation initiation factors, poly-A-binding protein, and small ribosomal subunits. Translation itself does not occur in stress granules, but mRNAs can become "translation-ready" as the proteins bound to them in P-bodies are replaced with those in stress granules. The movement of mRNAs between active translation, P-bodies, and stress granules can be seen as an mRNA cycle (**Figure 7–73**) where the competition between translation and mRNA degradation is carefully controlled. Thus, when translation initiation is blocked (by starvation, drugs, or genetic manipulation), stress granules enlarge as more and more nontranslated mRNAs are moved directly into them for storage. Clearly, once a cell has made the large investment in producing a properly processed mRNA molecule, it carefully controls its subsequent fate.

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.

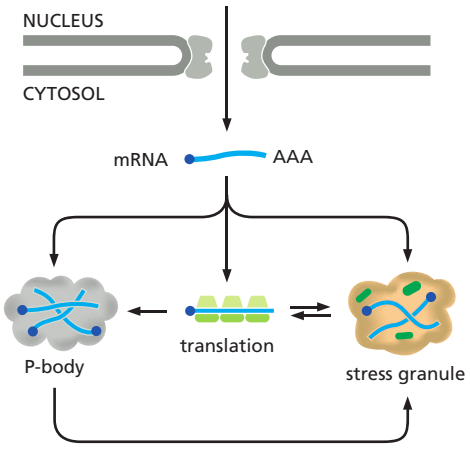


Figure 7–73 Possible fates of an mRNA molecule. An mRNA molecule released from the nucleus can be actively translated *(center)*, stored in stress granules *(right)*, or degraded in P-bodies *(left)*. As the needs of the cell change, mRNAs can be shuffled from one pool to the next, as indicated by the arrows.

REGULATION OF GENE EXPRESSION BY NONCODING RNAs

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 6–1). But we have seen throughout this book that RNA molecules perform many critical tasks in the cell besides serving as intermediate carriers of genetic information. Among these noncoding RNAs are the rRNA and tRNA molecules, which are responsible for reading the genetic code and synthesizing proteins. The RNA molecule in telomerase serves as a template for the replication of chromosome ends, snoRNAs modify ribosomal RNA, and snRNAs carry out the major events of RNA splicing. And we saw in the previous section that Xist RNA has an important role in inactivating one copy of the X chromosome in females.

A series of recent discoveries has revealed that noncoding RNAs are even more prevalent than previously imagined. We now know that such RNAs play widespread roles in regulating gene expression and in protecting the genome from viruses and transposable elements. These newly discovered RNAs are the subject of this section.

Small Noncoding RNA Transcripts Regulate Many Animal and Plant Genes Through RNA Interference

We begin our discussion with a group of short RNAs that carry out **RNA interference** or **RNAi**. Here, short single-stranded RNAs (20–30 nucleotides) serve as guide RNAs that selectively reorganize and bind—through base-pairing—other RNAs in the cell. When the target is a mature mRNA, the small noncoding RNAs can inhibit its translation or even catalyze its destruction. If the target RNA molecule is in the process of being transcribed, the small noncoding RNA can bind to it and direct the formation of certain types of repressive chromatin on its attached DNA template (**Figure 7–74**). Three classes of small noncoding RNAs work in this way—*microRNAs* (*miRNAs*), *small interfering RNAs* (*siRNAs*), and *piwi-interacting RNAs* (*piRNAs*)—and we discuss them in turn in the next sections. Although they differ in the way the short pieces of single-stranded RNA are generated, all three types of short RNAs locate their targets through RNA–RNA base-pairing, and they generally cause reductions in gene expression.

miRNAs Regulate mRNA Translation and Stability

Over 1000 different **microRNAs** (**miRNAs**) are produced from the human genome, and these appear to regulate at least one-third of all human protein-coding genes. Once made, miRNAs base-pair with specific mRNAs and fine-tune their translation and stability. The miRNA precursors are synthesized by RNA polymerase II and are capped and polyadenylated. They then undergo a special type of processing, after which the miRNA (typically 23 nucleotides in length) 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

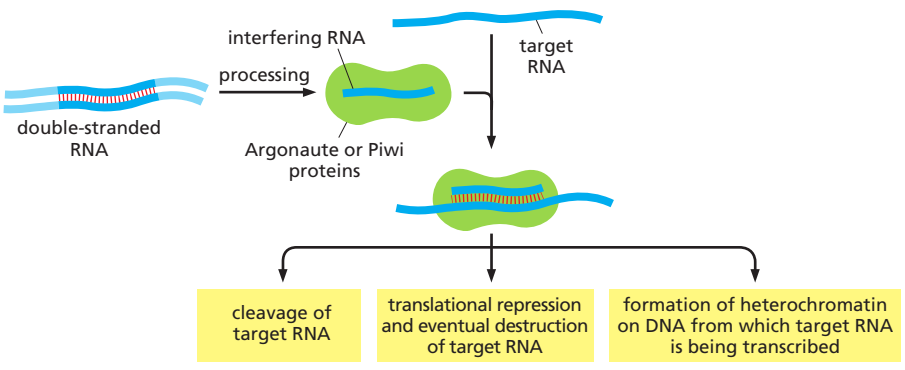


Figure 7–74 RNA interference in eukaryotes. Single-stranded interfering RNAs are generated from double-stranded RNA. They locate target RNAs through base-pairing and, at this point, several fates are possible, as shown. As described in the text, there are several types of RNA interference; the way the double-stranded RNA is produced and processed and the ultimate fate of the target RNA depends on the particular system.

sequences (Figure 7–75). This search is greatly facilitated by the Argonaute protein, a component of RISC, which holds the 5′ region of the miRNA so that it is optimally positioned for base-pairing to another RNA molecule (Figure 7–76). In animals, the extent of base-pairing is typically at least seven nucleotide pairs, and this pairing most often occurs 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 (which is unusual in humans but common in many plants), the mRNA is cleaved (*sliced*) by the Argonaute protein, effectively removing the mRNA's poly-A tail and exposing it to exonucleases (see Figure 7–69). Following cleavage of the mRNA, the RISC with its associated miRNA is released, and it can seek out additional mRNAs (see Figure 7–75). Thus, a single miRNA can act catalytically to destroy many complementary mRNAs. These miRNAs can be thus 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 (as observed for most human miRNAs), Argonaute does not slice the mRNA; rather, translation of the mRNA is repressed and the mRNA is shuttled to P-bodies (see Figure 7–73) where, sequestered from ribosomes, it eventually undergoes poly-A tail shortening, decapping, and degradation.

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 short sequence in their UTRs. This situation is common in humans, where a single miRNA can 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 same mRNA lead to further reductions in its translation. As discussed earlier for transcription regulators, 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

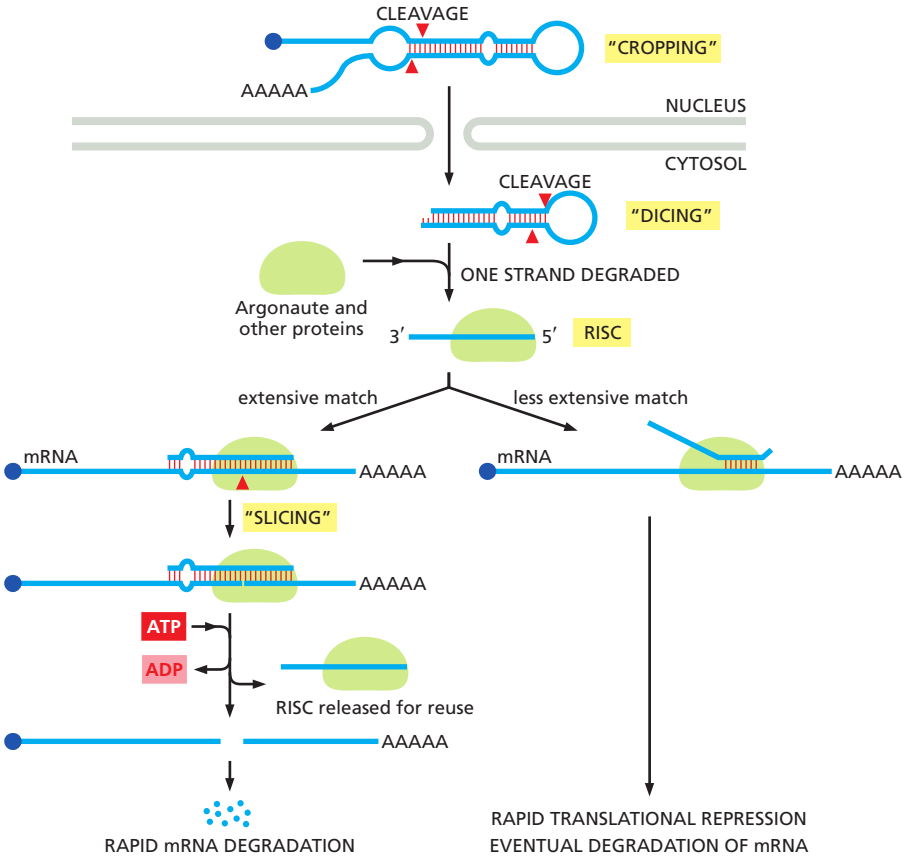


Figure 7–75 miRNA processing and mechanism of action. The precursor miRNA, through complementarity between one part of its sequence and another, forms a double-stranded structure. This RNA is cropped while still in the nucleus and then exported to the cytosol, where it is further cleaved by the Dicer enzyme to form the miRNA proper. Argonaute, in conjunction with other components of RISC, initially associates with both strands of the miRNA and then 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 mammals, the miRNA–mRNA match often does not extend beyond a short seven-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.

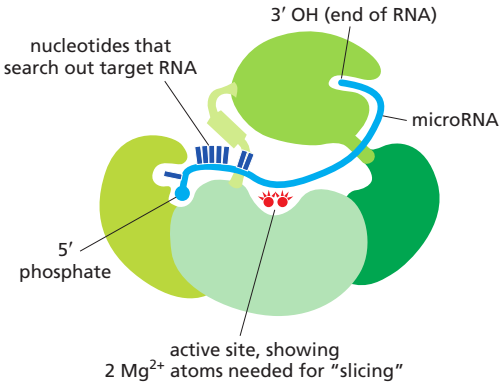


Figure 7–76 Human Argonaute protein carrying an miRNA. The protein is folded into four structural domains, each indicated by a different color. The miRNA is held in an extended form that is optimal for forming RNA–RNA base pairs. The active site of Argonaute that "slices" a target RNA, when it is extensively base-paired with the miRNA, is indicated in *red*. Many Argonaute proteins (three out of the four human proteins, for example) lack the catalytic site and therefore bind target RNAs without slicing them. (Adapted from C.D. Kuhn and L. Joshua-Tor, *Trends Biochem. Sci.* 38:263–271, 2013. With permission from Cell Press.)

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 appreciate the full impact of miRNAs, it is clear that they represent an important part of the cell's equipment for regulating the expression of genes. We discuss specific examples of miRNAs that have key roles in development in Chapter 21.

RNA Interference Is Also Used as 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 the degradation of foreign RNA molecules, specifically those that occur in double-stranded form. Many transposable elements and viruses produce double-stranded RNA, at least transiently, in their life cycles, and RNA interference helps to keep these potentially dangerous invaders in check. As we shall see, this form of RNAi also provides 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 7–75). This protein cleaves the double-stranded RNA into small fragments (approximately 23 nucleotide pairs) called **small interfering RNAs** (**siRNAs**). These double-stranded siRNAs are then bound by Argonaute and other components of RISC. As we saw above for miRNAs, one strand of the duplex RNA is then 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 usually exact, Argonaute cleaves these molecules, leading to their rapid destruction.

Each time RISC cleaves a new RNA molecule, the RISC 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 use siRNAs as primers to produce additional copies of double-strand RNAs which are then cleaved into siRNAs. 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 the specific 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 response resembles certain aspects of the animal immune system; in both, an invading organism elicits a customized response, and—through amplification of the "attack" molecules—the host becomes systemically protected.

We have seen that although miRNAs and siRNAs are generated in slightly different ways, they rely on the same proteins and seek out their targets in a fundamentally similar manner. Because siRNAs are found in widespread species, they are believed to be the most ancient form of RNA interference, with miRNAs being a later refinement. These siRNA-mediated defense mechanisms are crucial for plants, worms, and insects. In mammals, a protein-based system (described in Chapter 24) has largely taken over the task of fighting off viruses.

RNA Interference Can Direct Heterochromatin Formation

The siRNA interference pathway just described does not necessarily stop with the destruction of target RNA molecules. In some cases, the RNA interference machinery can also selectively shut off *synthesis* of the target RNAs. For this 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 (**Figure 7–77**). Positioned on the genome in this way, the RITS complex attracts proteins that covalently modify nearby histones and eventually direct the formation of heterochromatin to prevent further transcription initiation. In some cases, an RNA-dependent RNA polymerase and a Dicer enzyme are also recruited by the RITS complex to continually generate additional siRNAs *in situ*. This positive feedback loop ensures continued repression of the target gene even after the initiating siRNA molecules have disappeared.

RNAi-directed heterochromatin formation is an important cell defense mechanism that limits the spread of transposable elements in genomes by maintaining their DNA sequences in a transcriptionally silent form. However, this same mechanism is also used in some 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

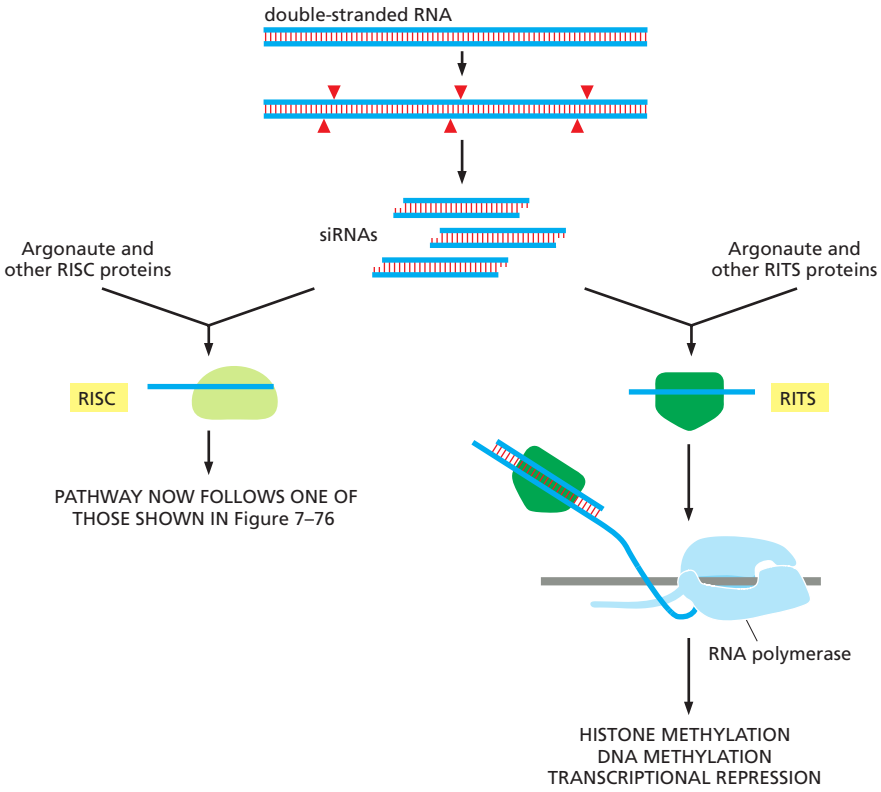


Figure 7–77 RNA interference directed by siRNAs. In many organisms, double-stranded RNA can trigger both the destruction of complementary mRNAs *(left)* and transcriptional silencing *(right)*. The change in chromatin structure induced by the bound RITS (RNA-induced transcriptional silencing) complex resembles that in Figure 7–45.

interference pathway and stimulates formation of the heterochromatin that surrounds centromeres, which is necessary for the centromeres to segregate chromosomes accurately during mitosis.

piRNAs Protect the Germ Line from Transposable Elements

A third system of RNA interference relies on **piRNAs** (**piwi-interacting RNAs**, named for Piwi, a class of proteins related to Argonaute). piRNAs are made specifically in the germ line, where they block the movement of transposable elements. Found in many organisms, including humans, genes coding for piRNAs consist largely of sequence fragments of transposable elements. These clusters of fragments are transcribed and broken up into short, single-stranded piRNAs. The processing differs from that for miRNAs and siRNAs (for one thing, the Dicer enzyme is not involved), and the resulting piRNAs are slightly longer than miRNAs and siRNAs; moreover, they are complexed with Piwi rather than Argonaute proteins. Once formed, the piRNAs seek out RNA targets by base-pairing and, much like siRNAs, transcriptionally silence intact transposon genes and destroy any RNA (including mRNAs) produced by them.

Many mysteries surround piRNAs. Over a million piRNA species are coded in the genomes of many mammals and expressed in the testes, yet only a small fraction seem to be directed against the transposons present in those genomes. Are the piRNAs remnants of past invaders? Do they cover so much "sequence space" that they are broadly protective for any foreign DNA? Another curious feature of piRNAs is that many of them (particularly if base-pairing does not have to be perfect) should, in principle, attack the normal mRNAs made by the organism, yet they do not. It has been proposed that these large numbers of piRNAs may form a system to distinguish "self" RNAs from "foreign" RNAs and attack only the latter. If this is the case, there must be a special way for the cell to spare its own RNAs. One idea is that RNAs produced in the previous generation of an organism are somehow registered and set aside from piRNA attack in subsequent generations. Whether or not this mechanism truly exists, and, if so, how it might work, are questions that demonstrate our incomplete understanding of the full implications of RNA interference.

RNA Interference Has Become a Powerful Experimental Tool

Although it likely arose as a defense mechanism against viruses and transposable elements, RNA interference, as we have seen, 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, which can be readily carried out in cultured cells and, in many cases, whole animals and plants, has made possible new genetic approaches in cell and molecular biology. We shall discuss it in detail in the following chapter where we cover modern genetic methods used to study cells (see pp. 499–501). RNAi also has 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. Although the mechanism of RNA interference was discovered a few decades ago, we are still being surprised by its mechanistic details and by its broad biological implications.

Bacteria Use Small Noncoding RNAs to Protect Themselves from Viruses

Bacteria make up the vast majority of the Earth's biomass and, not surprisingly, viruses that infect bacteria greatly outnumber plant and animal viruses. These viruses generally have DNA genomes. A recent discovery revealed that many species of bacteria (and almost all species of archaebacteria) use a repository of

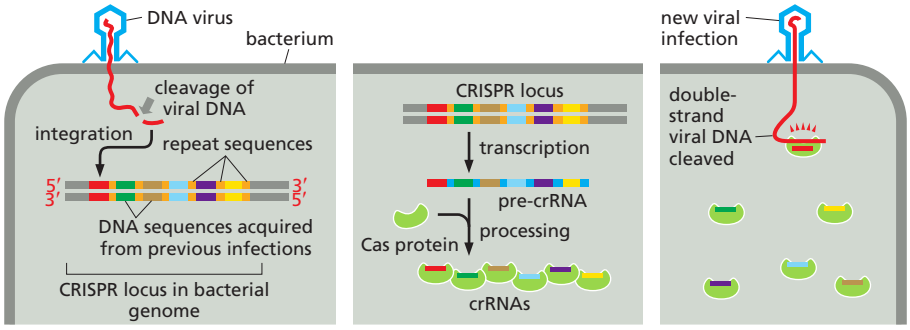
small noncoding RNA molecules to seek out and destroy the DNA of the invading viruses. Many features of this defense mechanism, known as the **CRISPR** system, resemble those we saw above for miRNAs and siRNAs, but there are two important differences. First, when bacteria and archaea are first infected by a virus, they have a mechanism that causes short fragments of that viral DNA to become integrated into their genomes. These serve as "vaccinations," in the sense that they become the templates for producing small noncoding RNAs known as **crRNAs** (CRISPR RNAs) that will thereafter destroy the virus should it reinfect the descendants of the original cell. This aspect of the CRISPR system is similar in principle to adaptive immunity in mammals, in that the cell carries a record of past exposures that is used to protect against future exposures. The second distinguishing feature of the CRISPR system is that these crRNAs then become associated with special proteins that allow them to seek out and destroy double-stranded DNA molecules, rather than single-stranded RNA molecules.

Although many details of CRISPR-mediated immunity remain to be discovered, we can outline the general process in three steps (**Figure 7–78**). In the first, viral DNA sequences are integrated into special regions of the bacterial genome known as CRISPR (clustered regularly interspersed short palindromic repeat) loci, named for the peculiar structure that first drew the attention of scientists. In its simplest form, a CRISPR locus consists of several hundred repeats of a host DNA sequence interspersed with a large collection of sequences (typically 25–70 nucleotide pairs each) that has been derived from prior exposures to viruses and other foreign DNA. The newest viral sequence is always integrated at the 5′ end of the CRISPR locus, the end that is transcribed first. Each locus, therefore, carries a temporal record of prior infections. Many bacterial and archaeal species carry several large CRISPR loci in their genomes and are thus immune to a wide range of viruses.

In the second step, the CRISPR locus is transcribed to produce a long RNA molecule, which is then processed into the much shorter (approximately 30 nucleotides) crRNAs. In the third step, crRNAs complexed with *Cas* (*CRISPR-associated*) proteins seek out complementary viral DNA sequences and direct their destruction by nucleases. Although structurally dissimilar, Cas proteins are analogous to the Argonaute and Piwi proteins discussed above: they hold small single-stranded RNAs in an extended configuration that is optimized, in this case, for seeking and forming complementary base pairs with DNA.

We still have much to learn about CRISPR-based immunity in bacteria and archaebacteria. The mechanism through which viral sequences are first identified and integrated into the host genome is poorly understood, as is the way that the crRNAs find their complementary sequences in double-stranded DNA. Moreover, in different species of bacteria and archaebacteria, crRNAs are processed in different ways, and in some cases, the crRNAs can attack viral RNAs as well as DNAs.

We shall see in the following chapter that bacterial CRISPR systems have already been artificially "moved" into plants and animals, where they have become very powerful experimental tools for manipulating genomes.



STEP 1: short viral DNA sequence is integrated into CRISPR locus

STEP 2: RNA is transcribed from CRISPR locus, processed, and bound to Cas protein

STEP 3: small crRNA in complex with Cas seeks out and destroys viral sequences

Figure 7–78 CRISPR-mediated immunity in bacteria and archaebacteria. After infection by a virus *(left panel)*, a small bit of DNA from the viral genome is inserted into the CRISPR locus. For this to happen, a small fraction of infected cells must survive the initial viral infection. The surviving cells, or more generally their descendants, transcribe the CRISPR locus and process the transcript into crRNAs *(middle panel)*. Upon reinfection with a virus that the population has already been "vaccinated" against, the incoming viral DNA is destroyed by a complementary crRNA *(right panel)*. For a CRISPR system to be effective, the crRNAs must not destroy the CRISPR locus itself, even though the crRNAs are complementary in sequence to it. In many species, in order for crRNAs to attack an invading DNA molecule, there must be additional short nucleotide sequences that are carried by the target molecule. Because these sequences, known as PAMs (protospacer adjacent motifs), lie outside the crRNA sequences, the host CRISPR locus is spared (see Figure 8–55).

Long Noncoding RNAs Have Diverse Functions in the Cell

In this and the preceding chapters, we have seen that noncoding RNA molecules have many functions in the cell. Yet, as is the case with proteins, there remain many noncoding RNAs whose function is still unknown. Many RNAs of unknown function belong to a group known as **long noncoding RNA** (**lncRNA**). These are arbitrarily defined as RNAs longer than 200 nucleotides that do not code for protein. As methods have improved for determining the nucleotide sequences of all the RNA molecules produced by a cell line or tissue, the sheer number of lncRNAs (an estimated 8000 for the human genome, for example) came as a surprise to scientists. Most lncRNAs are transcribed by RNA polymerase II and have 5′ caps and poly-A tails, and, in many cases, they are spliced. It has been difficult to annotate lncRNAs because low levels of RNA are now known to be made from 75% of the human genome. Most of these RNAs are thought to result from the background "noise" of transcription and RNA processing. According to this idea, such nonfunctional RNAs provide no fitness advantage or disadvantage to the organism and are a tolerated by-product of the complex patterns of gene expression that need to be produced in multicellular organisms. For these reasons, it is difficult to estimate the number of lncRNAs that are likely to have a function in the cell and to distinguish them from the background transcription.

We have already encountered a few lncRNAs, including the RNA in telomerase (see Figure 5–33), Xist RNA (see Figure 7–52), and an RNA involved in imprinting (see Figure 7–49). Other lncRNAs have been implicated in controlling the enzymatic activity of proteins, inactivating transcription regulators, affecting splicing patterns, and blocking translation of certain mRNAs.

In terms of biological function, lncRNA should be considered a catch-all phrase encompassing a great diversity of functions. Nevertheless, there are two unifying features of lncRNAs that can account for their many roles in the cell. The first is that lncRNAs can function as *scaffold RNA molecules*, holding together groups of proteins to coordinate their functions (**Figure 7–79**A). We have already seen an example in telomerase, where the RNA molecule holds together and organizes protein components. These RNA-based scaffolds are analogous to protein scaffolds we discussed in Chapter 3 (see Figure 3–78) and Chapter 6 (see Figure 6–47). RNA molecules are well suited to act as scaffolds: small bits of RNA sequence, often those portions that form stem-loop structures, can serve as binding sites for proteins, and these can be strung together with random sequences of RNA in between. This property may be one reason that lncRNAs show relatively little primary-sequence conservation across species.

The second key feature of lncRNAs is their ability to serve as guide sequences, binding to specific RNA or DNA target molecules through base-pairing. By doing so, they bring proteins that are bound to them into close proximity with the DNA

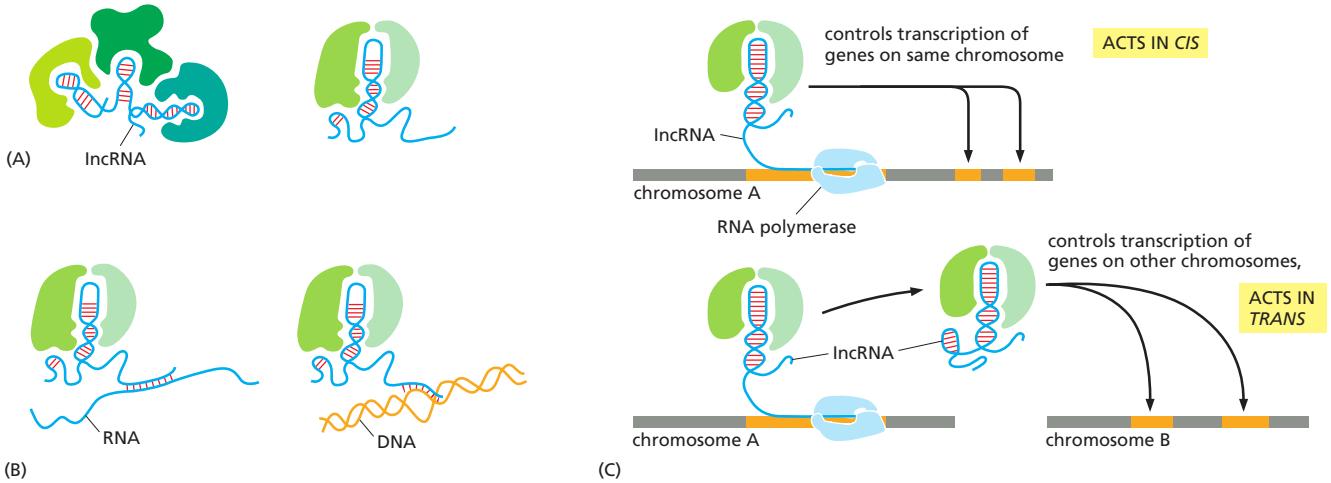


Figure 7–79 Roles of long noncoding RNA (lncRNA). (A) lncRNAs can serve as scaffolds, bringing together proteins that function in the same process. As described in Chapter 6, RNAs can fold into specific three-dimensional structures that are often recognized by proteins. (B) In addition to serving as scaffolds, lncRNAs can, through formation of complementary base pairs, localize proteins to specific sequences on RNA or DNA molecules. (C) In some cases, lncRNAs act only *in cis*, for example, when the RNA is held in place by RNA polymerase *(top)*. Other lncRNAs, however, diffuse from their sites of synthesis and therefore act *in trans*.

and RNA sequences (Figure 7–79B). This behavior is similar to that of snoRNAs (see Figure 6–41), crRNAs (see Figure 7–78), and miRNAs (see Figure 7–75), all of which act in this way to guide protein enzymes to specific nucleic acid sequences.

In some cases, lncRNAs work simply by base-pairing, without bringing in enzymes or other proteins. For example, a number of lncRNA genes are embedded in protein-coding genes, but they are transcribed in the "wrong direction." These *antisense RNAs* can form complementary base pairs with the mRNA (transcribed in the "correct" direction) and block its translation into protein (see Figure 7–66D). Other antisense lncRNAs base-pair with pre-mRNAs as they are synthesized and change the pattern of RNA splicing by masking splice-site sequences. Still others act as "sponges," base-pairing with miRNAs and thereby reducing their effects.

Finally, we note that some lncRNAs can act only *in cis*; that is, they affect only the chromosome from which they are transcribed. This readily occurs when the transcribed RNA has not yet been released from RNA polymerases (Figure 7–79C). Many lncRNAs, however, diffuse from their site of synthesis and act *in trans*. Although the best understood lncRNAs work in the nucleus, many are found in the cytosol. The functions—if any—of the great majority of these cytosolic lncRNAs remain undiscovered.

Summary

RNA molecules have many uses in the cell besides carrying the information needed to specify the order of amino acids during protein synthesis. Although we have encountered noncoding RNAs in other chapters (tRNAs, rRNAs, snoRNAs, for example), the sheer number of noncoding RNAs produced by cells has surprised scientists. One well understood use of noncoding RNAs occurs in RNA interference, where guide RNAs (miRNAs, siRNAs, piRNAs) 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 suppressing their transcription. In bacteria and archaebacteria, RNA interference is used as an adaptive immune response to destroy viruses that infect them. A large family of large noncoding RNAs (lncRNAs) has recently been discovered. Although the function of most of these RNAs is unknown, some serve as RNA scaffolds to bring specific proteins and RNA molecules together to speed up needed reactions.

WHAT WE DON'T KNOW

- How is the final rate of transcription of a gene specified by the hundreds of proteins that assemble on its control regions? Will we ever be able to predict this rate from inspection of the DNA sequences of control regions?
- How does the collection of *cis*-regulatory sequences embedded in a genome orchestrate the developmental program of a multicellular organism?
- How much of the human genome sequence is functional, and why is the remainder retained?
- Which of the thousands of unstudied noncoding RNAs have functions in the cell, and what are these functions?
- Were introns present in early cells (and subsequently lost in some organisms), or did they arise at later times?

PROBLEMS

Which statements are true? Explain why or why not.

7–1 In terms of the way it interacts with DNA, 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 Once cells have differentiated to their final specialized forms, they never again alter expression of their genes.

7–3 CG islands are thought to have arisen during evolution because they were associated with portions of the genome that remained unmethylated in the germ line.

7–4 In most differentiated tissues, daughter cells retain a memory of gene expression patterns that were present in the parent cell through mechanisms that do not involve changes in the sequence of their genomic DNA.

Discuss the following problems.

7–5 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

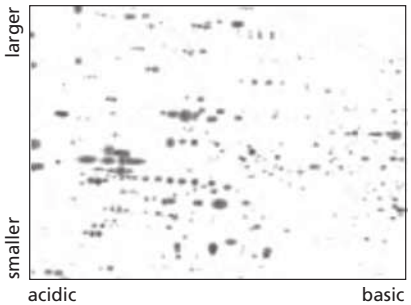


Figure Q7–1 Two-dimensional separation of proteins from the human brain (Problem 7–5). The proteins were displayed using two-dimensional gel electrophoresis. Only a small portion of the protein spectrum is shown. (Courtesy of Tim Myers and Leigh Anderson, Large Scale Biology Corporation.)

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.

7–6 Comparisons of the patterns of mRNA levels across different human cell types show 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?

7–7 What are the two fundamental components of a genetic switch?

7–8 The nucleus of a eukaryotic cell is much larger than a bacterium, and it contains much more DNA. As a consequence, a transcription regulator in a eukaryotic cell must be able to select its specific binding site from among many more unrelated sequences than does a transcription regulator in a bacterium. Does this present any special problems for eukaryotic gene regulation?

Consider the following situation. Assume that the eukaryotic 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 the concentration of the transcription regulator that binds the site were the same in the nucleus and in the bacterium, would the regulator occupy its binding site equally as well in the eukaryotic nucleus as it does in the bacterium? Explain your answer.

7–9 Some transcription regulators bind to DNA and cause the double helix to bend at a sharp angle. Such "bending proteins" can affect the initiation of transcription without directly contacting any other protein. Can you devise a plausible explanation for how such proteins might work to modulate transcription? Draw a diagram that illustrates your explanation.

7–10 How is it that protein–protein interactions that are too weak to cause proteins to assemble in solution can nevertheless allow the same proteins to assemble into complexes on DNA?

7–11 Imagine the two situations shown in **Figure Q7–2**. In cell 1, a transient signal induces the synthesis of protein A, which is a transcription activator that turns on many genes including its own. In cell 2, a transient signal induces the synthesis of protein R, which is a transcription repressor that turns off many genes including its own. In which, if either, of these situations will the descendants of the original cell "remember" that the progenitor cell had experienced the transient signal? Explain your reasoning.

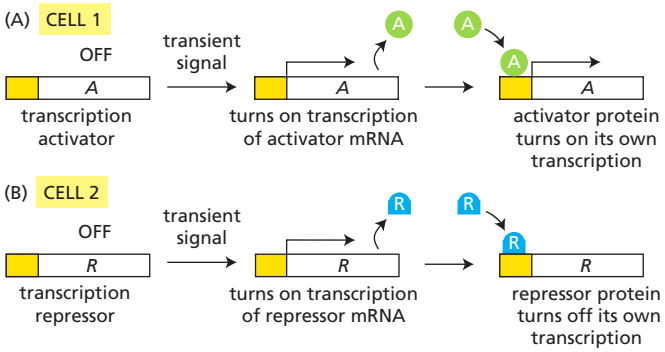


Figure Q7–2 Gene regulatory circuits and cell memory (Problem 7–11). (A) Induction of synthesis of transcription activator A by a transient signal. (B) Induction of synthesis of transcription repressor R by a transient signal.

7–12 Examine the two pedigrees shown in **Figure Q7–3**. One results from deletion of a maternally imprinted autosomal gene. The other pedigree results from deletion of a paternally imprinted autosomal gene. In both pedigrees, affected individuals (*red* symbols) are heterozygous for the deletion. These individuals are affected because one copy of the chromosome carries an imprinted, inactive gene, while the other carries a deletion of the gene. Dotted *yellow* symbols indicate individuals that carry the deleted locus, but do not display the mutant phenotype. Which pedigree is based on paternal imprinting and which on maternal imprinting? Explain your answer.

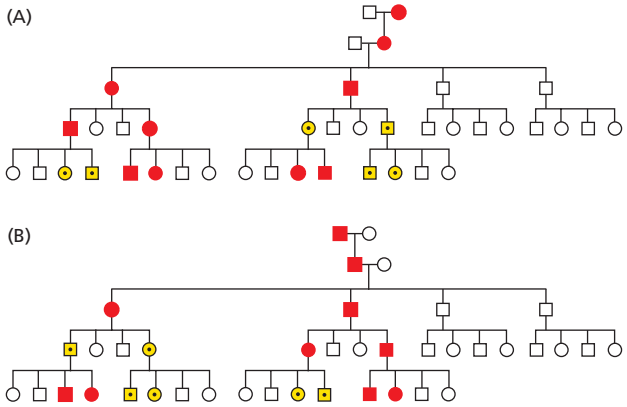


Figure Q7–3 Pedigrees reflecting maternal and paternal imprinting (Problem 7–12). In one pedigree, the gene is paternally imprinted; in the other, it is maternally imprinted. In generations 3 and 4, only one of the two parents in the indicated matings is shown; the other parent is a normal individual from outside this pedigree. Affected individuals are represented by *red circles* for females and *red squares* for males. *Dotted yellow symbols* indicate individuals that carry the deletion but do not display the phenotype.

7–13 If you insert a *β-galactosidase* gene lacking its own transcription control region into a cluster of piRNA genes in *Drosophila*, you find that β-galactosidase expression from a normal copy elsewhere in the genome is strongly inhibited in the fly's germ cells. If the inactive *β-galactosidase* gene is inserted outside the piRNA gene cluster, the normal gene is properly expressed. What do you suppose is the basis for this observation? How would you test your hypothesis?

REFERENCES

General

Brown TA (2007) Genomes 3. New York: Garland Science.

Epigenetics (2004) *Cold Spring Harb. Symp. Quant. Biol.* 69.

Gilbert SF (2013) Developmental Biology, 10th ed. Sunderland, MA: Sinauer Associates, Inc.

Hartwell L, Hood L, Goldberg ML et al. (2010) Genetics: from Genes to Genomes, 4th ed. Boston: McGraw Hill.

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, NY: Cold Spring Harbor Laboratory Press.

Watson J, Baker T, Bell S et al. (2013) Molecular Biology of the Gene, 7th ed. Menlo Park, CA: Benjamin Cummings.

An Overview of Gene Control

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.

Kellis M, Wold B, Synder MP et al. (2014) Defining functional DNA elements in the human genome. *Proc. Natl. Acad. Sci. USA* 111, 6131–6138.

Control Of Transcription By Sequence-Specific DNA-Binding Proteins

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.

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.

Weirauch MT & Hughes TR (2011) A catalogue of eukaryotic transcription factor types, their evolutionary origin, and species distribution. In A Handbook of Transcription Factors. New York, NY: Springer Publishing Company.

Transcription Regulators Switch Genes On and Off

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.

Gilbert W & Müller-Hill B (1967) The lac operator is DNA. *Proc. Natl. Acad. Sci. USA* 58, 2415.

Jacob F & Monod J (1961) Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.* 3, 318–356.

Levine M, Cattoglio C & Tjian R (2014) Looping back to leap forward: transcription enters a new era. *Cell* 157, 13–25.

Narlikar GJ, Sundaramoorthy R & Owen-Hughes T (2013) Mechanisms and Functions of ATP-dependent chromatin-remodeling enzymes. *Cell* 154, 490–503.

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.

Turner BM (2014) Nucleosome signaling: an evolving concept. *Biochim. Biophys. Acta* 1839, 623–626.

Molecular Genetic Mechanisms that Create and Maintain Specialized Cell Types

Alon U (2007) Network motifs: theory and experimental approaches. *Nature* 8, 450–461.

Buganim Y, Faddah DA & Jaenisch R (2013) Mechanisms and models of somatic cell reprogramming. *Nat. Rev. Genet.* 14, 427–439.

Hobert O (2011) Regulation of Terminal differentiation programs in the nervous system. *Annu. Rev. Cell Dev. Biol.* 27, 681–696.

Lawrence PA (1992) The Making of a Fly: The Genetics of Animal Design. New York: Blackwell Scientific Publications.

Mechanisms That Reinforce Cell Memory in Plants and Animals

Bird A (2011) Putting the DNA back into DNA methylation. *Nat. Genet.* 43, 1050–1051.

Gehring M (2013) Genomic imprinting: insights from plants. *Annu. Rev. Genet.* 47, 187–208.

Lawson HA, Cheverud JM & Wolf JB (2013) Genomic imprinting and parent-of-origin effects on complex traits. *Genetics* 14, 609–617.

Lee JT & Bartolomei MS (2013) X-Inactivation, imprinting, and long noncoding RNAs in Health and disease. *Cell* 152, 1308–1323.

Li E & Zhang Y (2014) DNA methylation in mammals. *Cold Spring Harb. Perspect. Biol.* 6, a019133.

Post-Transcriptional Controls

DiGiammartino DC, Nishida K & Manley JL (2011) Mechanisms and consequences of alternative polyadenylation. *Mol Cell* 43, 853–866.

Gottesman S & Storz G (2011) Bacterial small RNA regulators: versatile roles and rapidly evolving variations. *Cold Spring Harb. Perspect. Biol.* 3, a003798.

Hershey JWB, Sonenberg N & Mathews MB (2012) Principles of translational control: an overview. *Cold Spring Harb. Perspect. Biol.* 4, a011528.

Hundley HA & Bass BL (2010) ADAR editing in double-stranded UTRs and other noncoding RNA sequences. *Trends Biochem. Sci.* 35, 377–383.

Kalsotra A & Cooper TA (2011) Functional consequences of developmentally regulated alternative splicing. *Nat. Rev. Genet.* 12, 715–729.

Kortmann J & Narberhaus F (2012) Bacterial RNA thermometers: molecular zippers and switches. *Nat. Rev. Microbiol.* 10, 255–265.

Parker R (2012) RNA degradation in *Saccharomyces cerevisae*. *Genetics* 191, 671–702.

Popp MW & Maquat LE (2013) Organizing principles of mammalian nonsense-mediated mRNA decay. *Annu. Rev. Genet.* 47, 139–165.

Serganov A & Nudler E (2013) A decade of riboswitches. *Cell* 152, 17–24.

Thompson SR (2012) Tricks an IRES uses to enslave ribosomes. *Trends Microbiol.* 20, 558–566.

Regulation of Gene Expression By Noncoding RNAs

Bhaya D, Davison M & Barrangou R (2011) CRISPR-Cas systems in bacteria and archaea: Versatile small RNAs for adaptive defense and regulation. *Annu. Rev. Genet.* 45, 273–297.

Cech TR & Steitz JA (2014) The noncoding RNA revolution–trashing old rules to forge new ones. *Cell* 157, 77–94.

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.

Guttman M & Rinn JL (2012) Modular regulatory principles of large non-coding RNAs. *Nature* 482, 339–346.

Lee HC, Gu W, Shirayama M et al. (2012) *C. elegans* piRNAs mediate the genome-wide surveillance of germline transcripts. *Cell* 150, 78–87.

Meister G (2013) Argonaute proteins: functional insights and emerging roles. *Nat. Rev. Genet.* 14, 447–459.

Rinn JL & Chang HY (2012) Genome regulation by long noncoding RNAs. *Annu. Rev. Biochem.* 81, 145–166.

tenOever BR (2013) RNA viruses and the host microRNA machinery. *Nat. Rev. Microbiol.* 11, 169–180.

Ulitsky I & Bartel DP (2013) lincRNAs: genomics, evolution, and mechanisms. *Cell* 154, 26–46.

Wiedenheft B, Sternberg SH & Doudna JA (2012) RNA-guided genetic silencing systems in bacteria and archaea. *Nature* 482, 331–338.