How Cells Read the Genome: From DNA to Protein

6

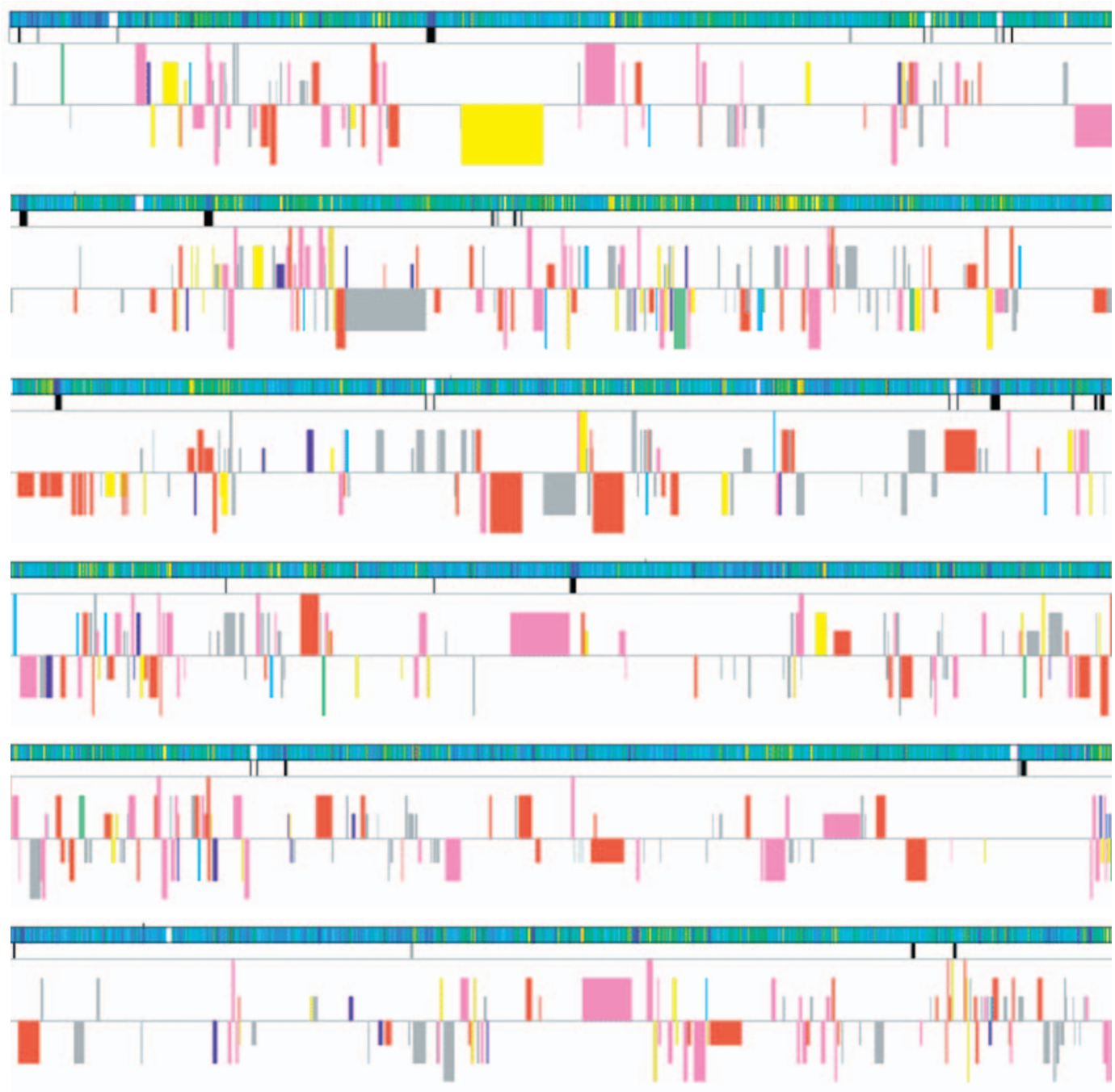
In This Chapter

Only when the structure of DNA was discovered in the early 1950s did it become clear how the hereditary information in cells is encoded in DNA's sequence of nucleotides. The progress since then has been astounding. Within fifty years we knew the complete genome sequences for many organisms, including humans. We therefore know the maximum amount of information that is required to produce a complex organism like ourselves. The limits on the hereditary information needed for life constrain the biochemical and structural features of cells and make it clear that biology is not infinitely complex.

In this chapter, we explain how cells decode and use the information in their genomes. Much has been learned about how the genetic instructions written in an alphabet of just four "letters"—the four different nucleotides in DNA—direct the formation of a bacterium, a fruit fly, or a human. Nevertheless, we still have a great deal to discover about how the information stored in an organism's genome produces even the simplest unicellular bacterium with 500 genes, let alone how it directs the development of a human with approximately 25,000 genes. An enormous amount of ignorance remains; many fascinating challenges therefore await the next generation of cell biologists.

The problems that cells face in decoding genomes can be appreciated by considering a small portion of the genome of the fruit fly *Drosophila melanogaster* (**Figure 6–1**). Much of the DNA-encoded information present in this and other genomes specifies the linear order—the sequence—of amino acids for every protein the organism makes. As described in Chapter 3, the amino acid sequence in turn dictates how each protein folds to give a molecule with a distinctive shape and chemistry. When a cell makes a particular protein, it must decode accurately the corresponding region of the genome. Additional information encoded in the DNA of the genome specifies exactly when in the life of an organism and in which cell types each gene is to be expressed into protein. Since proteins are the main constituents of cells, the decoding of the genome determines not only the size, shape, biochemical properties, and behavior of cells, but also the distinctive features of each species on Earth.

One might have predicted that the information present in genomes would be arranged in an orderly fashion, resembling a dictionary or a telephone directory. Although the genomes of some bacteria seem fairly well organized, the genomes of most multicellular organisms, such as our *Drosophila* example, are surprisingly disorderly. Small bits of coding DNA (that is, DNA that codes for protein) are interspersed with large blocks of seemingly meaningless DNA. Some sections of the genome contain many genes and others lack genes altogether. Proteins that work closely with one another in the cell often have their genes located on different chromosomes, and adjacent genes typically encode proteins that have little to do with each other in the cell. Decoding genomes is therefore no simple matter. Even with the aid of powerful computers, it is still difficult for researchers to locate definitively the beginning and end of genes in the DNA sequences of complex genomes, much less to predict when each gene is expressed in the life of the organism. Although the DNA sequence of the human genome is known, it will probably take at least a decade to identify every gene and determine the precise amino acid sequence of the protein it produces. Yet the cells in our body do this thousands of times a second.



KEY:

13 or more
1 to 12
none

length of bar indicates number of corresponding cDNAs identified in databases

100,000 nucleotide pairs

%GC content 25 65

transposable elements

known and predicted genes identified on top strand of DNA

known and predicted genes identified on bottom strand of DNA

color code for sequence similarity of genes identified

MWY
WY
MW
W
MY
Y
M
no similarity to MWY

M = mammalian
W = *C. elegans*
Y = *S. cerevisiae*

Figure 6–1 ***(opposite page)*** **Schematic depiction of a portion of chromosome 2 from the genome of the fruit fly *Drosophila melanogaster*.** This figure represents approximately 3% of the total *Drosophila* genome, arranged as six contiguous segments. As summarized in the key, the symbolic representations are: *black vertical lines* of various thicknesses: locations of transposable elements, with thicker bars indicating clusters of elements; *colored boxes*: genes (both known and predicted) coded on one strand of DNA (boxes *above* the midline) and genes coded on the other strand (boxes *below* the midline). The length of each gene box includes both its exons (protein-coding DNA) and its introns (noncoding DNA) (see Figure 4–15); its height is proportional to the number of known cDNAs that match the gene. (As described in Chapter 8, cDNAs are DNA copies of mRNA molecules, and large collections of the nucleotide sequences of cDNAs have been deposited in a variety of databases, the more matches, the higher the confidence that the predicted gene is transcribed into RNA and is thus a genuine gene.) The color of each gene box indicates whether a closely related gene is known to occur in other organisms. For example, MWY means the gene has close relatives in mammals, in the nematode worm *Caenorhabditis elegans*, and in the yeast *Saccharomyces cerevisiae*. MW indicates the gene has close relatives in mammals and the worm but not in yeast. The *rainbow-colored bar* indicates percent G–C base pairs; across many different genomes, this percentage shows a striking regional variation, whose origin and significance are uncertain. (From M.D. Adams et al., *Science* 287:2185–2195, 2000. With permission from AAAS.)

The DNA in genomes does not direct protein synthesis itself, but instead uses RNA as an intermediary. When the cell needs a particular protein, the nucleotide sequence of the appropriate portion of the immensely long DNA molecule in a chromosome is first copied into RNA (a process called *transcription)*. It is these RNA copies of segments of the DNA that are used directly as templates to direct the synthesis of the protein (a process called *translation)*. The flow of genetic information in cells is therefore from DNA to RNA to protein (**Figure 6–2**). All cells, from bacteria to humans, express their genetic information in this way—a principle so fundamental that it is termed the *central dogma* of molecular biology.

Despite the universality of the central dogma, there are important variations in the way in which information flows from DNA to protein. Principal among these is that RNA transcripts in eucaryotic cells are subject to a series of processing steps in the nucleus, including *RNA splicing*, before they are permitted to exit from the nucleus and be translated into protein. These processing steps can critically change the "meaning" of an RNA molecule and are therefore crucial for understanding how eucaryotic cells read their genomes. Finally, although we focus on the production of the proteins encoded by the genome in this chapter, we see that for many genes RNA is the final product. Like proteins, many of these RNAs fold into precise three-dimensional structures that have structural, catalytic, and regulatory roles in the cell.

We begin this chapter with the first step in decoding a genome: the process of transcription by which an RNA molecule is produced from the DNA of a gene. We then follow the fate of this RNA molecule through the cell, finishing when a correctly folded protein molecule has been formed. At the end of the chapter, we consider how the present quite complex scheme of information storage, transcription, and translation might have arisen from simpler systems in the earliest stages of cell evolution.

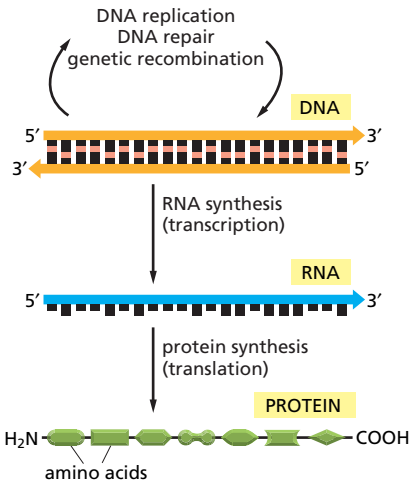


Figure 6–2 The pathway from DNA to protein. The flow of genetic information from DNA to RNA (transcription) and from RNA to protein (translation) occurs in all living cells.

FROM DNA TO RNA

Transcription and translation are the means by which cells read out, or express, the genetic instructions in their genes. Because many identical RNA copies can be made from the same gene, and each RNA molecule can direct the synthesis of many identical protein molecules, cells can synthesize a large amount of protein rapidly when necessary. But each gene can also be transcribed and translated with a different efficiency, allowing the cell to make vast quantities of some proteins and tiny quantities of others (**Figure 6–3**). Moreover, as we see in the next chapter, a cell can change (or regulate) the expression of each of its genes according to the needs of the moment—most commonly by controlling the production of its RNA.

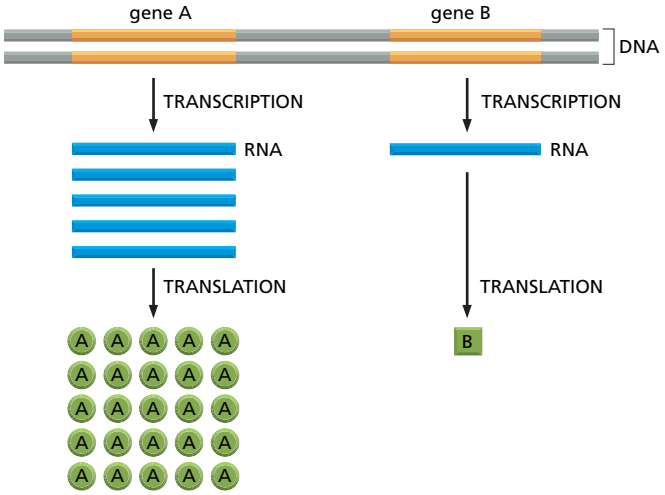


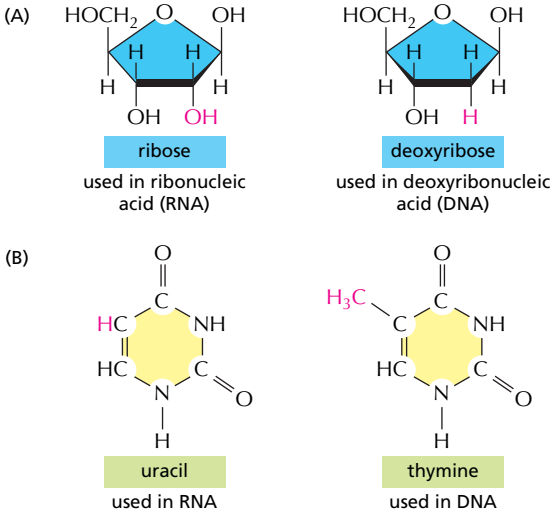
Figure 6–3 Genes can be expressed with different efficiencies. In this example, gene A is transcribed and translated much more efficiently than gene B. This allows the amount of protein A in the cell to be much greater than that of protein B.

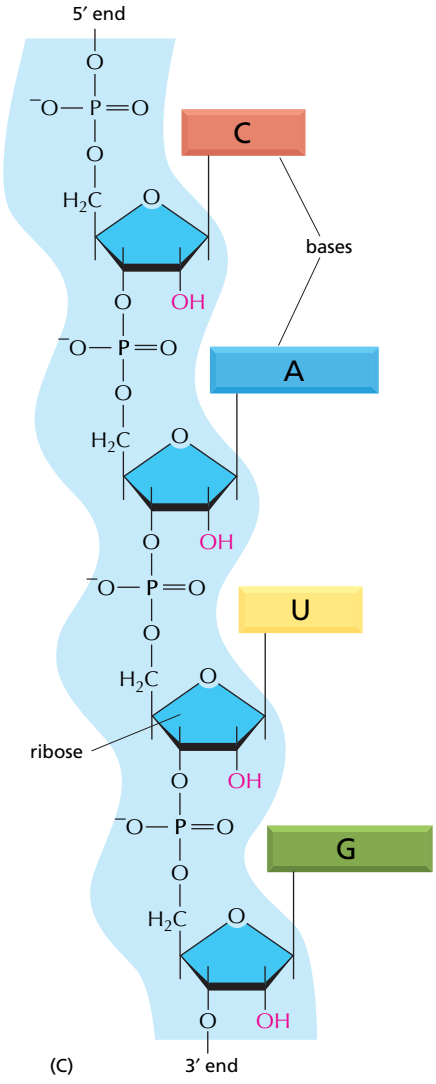
Portions of DNA Sequence Are Transcribed into RNA

The first step a cell takes in reading out a needed part of its genetic instructions is to copy a particular portion of its DNA nucleotide sequence—a gene—into an RNA nucleotide sequence. The information in RNA, although copied into another chemical form, is still written in essentially the same language as it is in DNA—the language of a nucleotide sequence. Hence the name **transcription**.

Like DNA, RNA is a linear polymer made of four different types of nucleotide subunits linked together by phosphodiester bonds (**Figure 6–4**). It differs from DNA chemically in two respects: (1) the nucleotides in RNA are *ribonucleotides*—that is, they contain the sugar ribose (hence the name *ribo*nucleic acid) rather than deoxyribose; (2) although, like DNA, RNA contains the bases adenine (A), guanine (G), and cytosine (C), it contains the base uracil (U) instead of the thymine (T) in DNA. Since U, like T, can base-pair by hydrogen-bonding with A (**Figure 6–5**), the complementary base-pairing properties described for DNA in Chapters 4 and 5 apply also to RNA (in RNA, G pairs with C, and A pairs with U). We also find other types of base pairs in RNA: for example, G occasionally pairs with U.

Figure 6–4 The chemical structure of RNA. (A) RNA contains the sugar ribose, which differs from deoxyribose, the sugar used in DNA, by the presence of an additional –OH group. (B) RNA contains the base uracil, which differs from thymine, the equivalent base in DNA, by the absence of a $-CH_3$ group. (C) A short length of RNA. The phosphodiester chemical linkage between nucleotides in RNA is the same as that in DNA.





Although these chemical differences are slight, DNA and RNA differ quite dramatically in overall structure. Whereas DNA always occurs in cells as a double-stranded helix, RNA is single-stranded. An RNA chain can therefore fold up into a particular shape, just as a polypeptide chain folds up to form the final shape of a protein (**Figure 6–6**). As we see later in this chapter, the ability to fold into complex three-dimensional shapes allows some RNA molecules to have precise structural and catalytic functions.

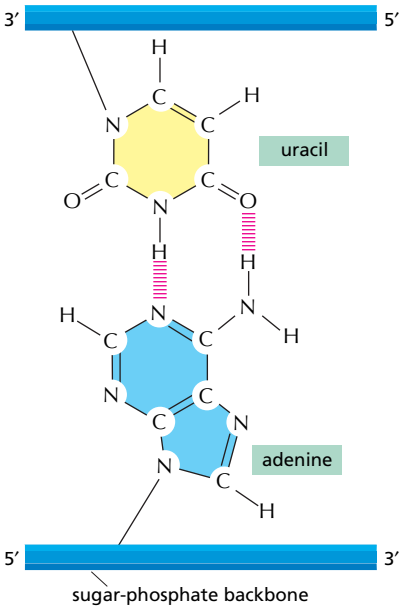


Figure 6–5 Uracil forms base pairs with adenine. The absence of a methyl group in U has no effect on base-pairing; thus, U–A base pairs closely resemble T–A base pairs (see Figure 4–4).

Transcription Produces RNA Complementary to One Strand of DNA

The RNA in a cell is made by DNA transcription, a process that has certain similarities to the process of DNA replication discussed in Chapter 5. Transcription begins with the opening and unwinding of a small portion of the DNA double helix to expose the bases on each DNA strand. One of the two strands of the DNA double helix then acts as a template for the synthesis of an RNA molecule. As in DNA replication, the nucleotide sequence of the RNA chain is determined by the complementary base-pairing between incoming nucleotides and the DNA template. When a good match is made, the incoming ribonucleotide is covalently linked to the growing RNA chain in an enzymatically catalyzed reaction. The RNA chain produced by transcription—the *transcript*—is therefore elongated one nucleotide at a time, and it has a nucleotide sequence that is exactly complementary to the strand of DNA used as the template (**Figure 6–7**).

Transcription, however, differs from DNA replication in several crucial ways. Unlike a newly formed DNA strand, the RNA strand does not remain hydrogen-bonded to the DNA template strand. Instead, just behind the region where the ribonucleotides are being added, the RNA chain is displaced and the DNA helix re-forms. Thus, the RNA molecules produced by transcription are released from the DNA template as single strands. In addition, because they are copied from only a limited region of the DNA, RNA molecules are much shorter than DNA molecules. A DNA molecule in a human chromosome can be up to 250 million nucleotide-pairs long; in contrast, most RNAs are no more than a few thousand nucleotides long, and many are considerably shorter.

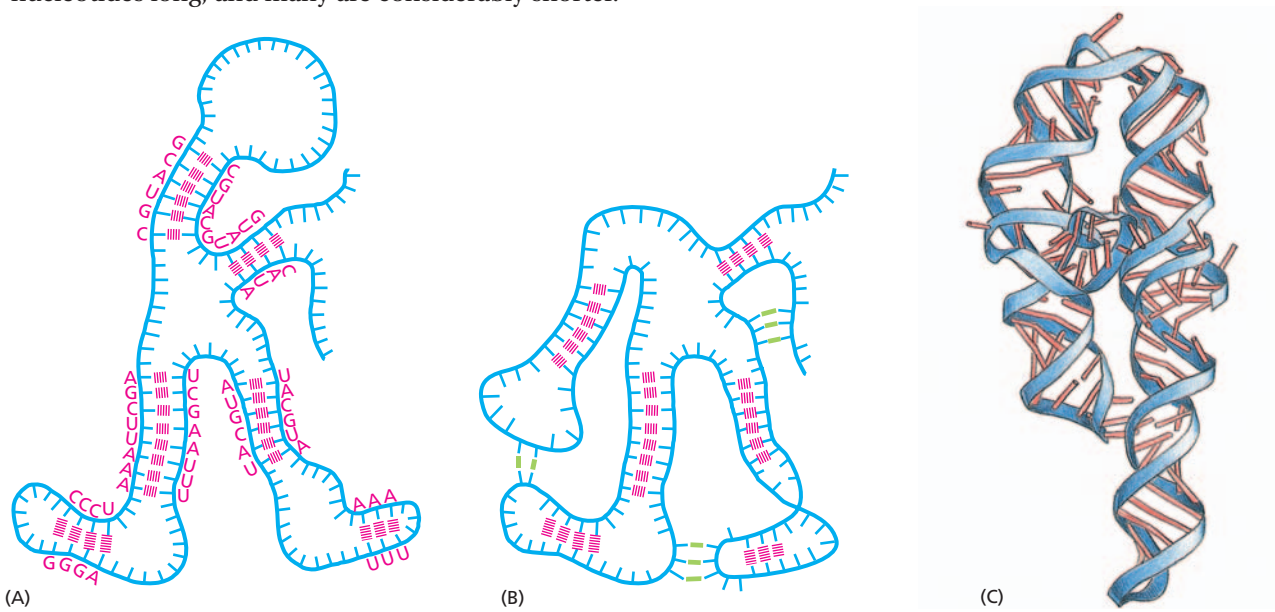


Figure 6–6 RNA can fold into specific structures. RNA is largely single-stranded, but it often contains short stretches of nucleotides that can form conventional base pairs with complementary sequences found elsewhere on the same molecule. These interactions, along with additional "nonconventional" base-pair interactions, allow an RNA molecule to fold into a three-dimensional structure that is determined by its sequence of nucleotides. <AATC> (A) Diagram of a folded RNA structure showing only conventional base-pair interactions. (B) Structure with both conventional *(red)* and nonconventional *(green)* base-pair interactions. (C) Structure of an actual RNA, a portion of a group I intron (see Figure 6–36). Each conventional base-pair interaction is indicated by a "rung" in the double helix. Bases in other configurations are indicated by broken rungs.

The enzymes that perform transcription are called **RNA polymerases**. Like the DNA polymerase that catalyzes DNA replication (discussed in Chapter 5), RNA polymerases catalyze the formation of the phosphodiester bonds that link the nucleotides together to form a linear chain. The RNA polymerase moves stepwise along the DNA, unwinding the DNA helix just ahead of the active site for polymerization to expose a new region of the template strand for complementary base-pairing. In this way, the growing RNA chain is extended by one nucleotide at a time in the 5′-to-3′ direction (**Figure 6–8**). The substrates are nucleoside triphosphates (ATP, CTP, UTP, and GTP); as in DNA replication, the hydrolysis of high-energy bonds provides the energy needed to drive the reaction forward (see Figure 5–4).

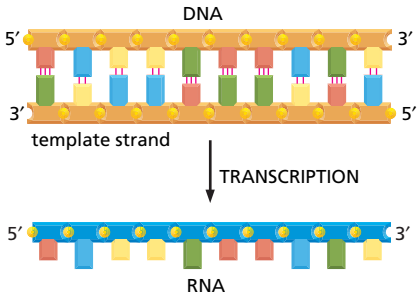


Figure 6–7 DNA transcription produces a single-stranded RNA molecule that is complementary to one strand of DNA.

The almost immediate release of the RNA strand from the DNA as it is synthesized means that many RNA copies can be made from the same gene in a relatively short time, with the synthesis of additional RNA molecules being started before the first RNA is completed (**Figure 6–9**). When RNA polymerase molecules follow hard on each other's heels in this way, each moving at about 20 nucleotides per second (the speed in eucaryotes), over a thousand transcripts can be synthesized in an hour from a single gene.

Although RNA polymerase catalyzes essentially the same chemical reaction as DNA polymerase, there are some important differences between the activities of the two enzymes. First, and most obviously, RNA polymerase catalyzes the linkage of ribonucleotides, not deoxyribonucleotides. Second, unlike the DNA polymerases involved in DNA replication, RNA polymerases can start an RNA chain without a primer. This difference may exist because transcription need not be as accurate as DNA replication (see Table 5–1, p. 271). Unlike DNA, RNA does not permanently store genetic information in cells. RNA polymerases make about one mistake for every 10^4 nucleotides copied into RNA (compared with an error rate for direct copying by DNA polymerase of about one in 10^7 nucleotides), and the consequences of an error in RNA transcription are much less significant than that in DNA replication.

Although RNA polymerases are not nearly as accurate as the DNA polymerases that replicate DNA, they nonetheless have a modest proofreading mechanism. If an incorrect ribonucleotide is added to the growing RNA chain, the polymerase can back up, and the active site of the enzyme can perform an excision reaction that resembles the reverse of the polymerization reaction,

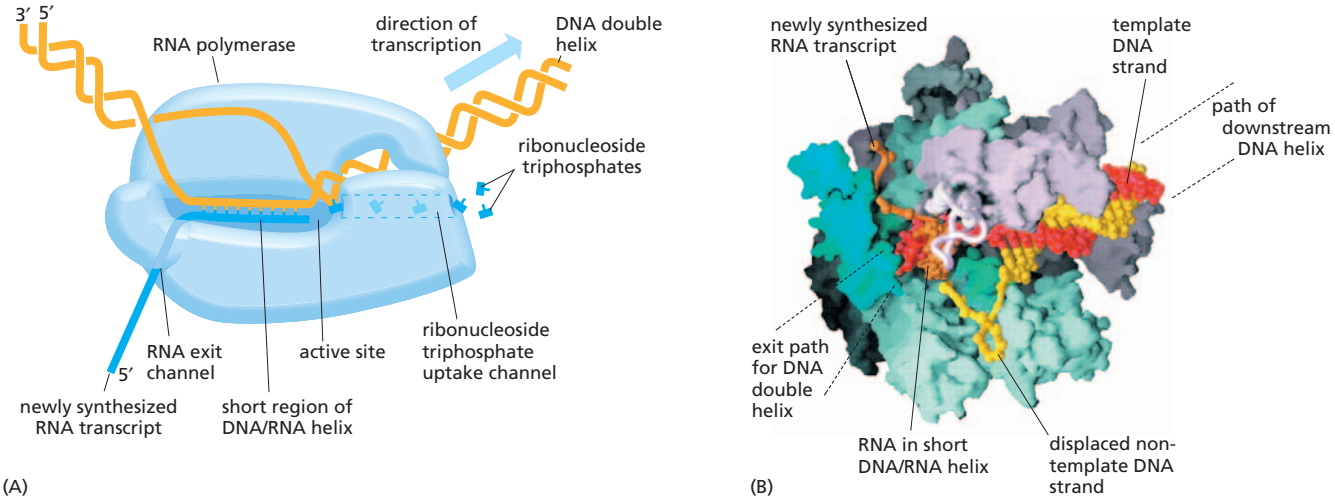


Figure 6–8 DNA is transcribed by the enzyme RNA polymerase. (A) The RNA polymerase *(pale blue)* moves stepwise along the DNA, unwinding the DNA helix at its active site. As it progresses, the polymerase adds nucleotides (represented as *small "T" shapes)* one by one to the RNA chain at the polymerization site, using an exposed DNA strand as a template. The RNA transcript is thus a complementary copy of one of the two DNA strands. A short region of DNA/RNA helix (approximately nine nucleotide pairs in length) is therefore formed only transiently, and a "window" of DNA/RNA helix therefore moves along the DNA with the polymerase. The incoming nucleotides are in the form of ribonucleoside triphosphates (ATP, UTP, CTP, and GTP), and the energy stored in their phosphate–phosphate bonds provides the driving force for the polymerization reaction (see Figure 5–4). (B) The structure of a bacterial RNA polymerase, as determined by x-ray crystallography. Four different subunits, indicated by different colors, comprise this RNA polymerase. The DNA strand used as a template is *red*, and the nontemplate strand is *yellow*. (A, adapted from a figure courtesy of Robert Landick; B, courtesy of Seth Darst.)

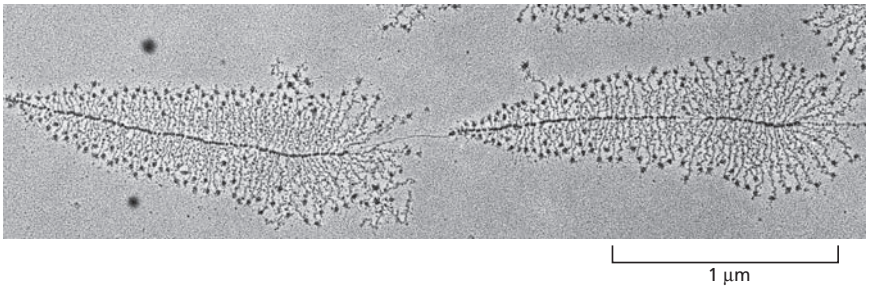


Figure 6–9 Transcription of two genes as observed under the electron microscope. The micrograph shows many molecules of RNA polymerase simultaneously transcribing each of two adjacent genes. Molecules of RNA polymerase are visible as a series of dots along the DNA with the newly synthesized transcripts (fine threads) attached to them. The RNA molecules (ribosomal RNAs) shown in this example are not translated into protein but are instead used directly as components of ribosomes, the machines on which translation takes place. The particles at the 5′ end (the free end) of each rRNA transcript are believed to reflect the beginnings of ribosome assembly. From the lengths of the newly synthesized transcripts, it can be deduced that the RNA polymerase molecules are transcribing from left to right. (Courtesy of Ulrich Scheer.)

except that water instead of pyrophosphate is used and a nucleoside monophosphate is released.

Given that DNA and RNA polymerases both carry out template-dependent nucleotide polymerization, it might be expected that the two types of enzymes would be structurally related. However, x-ray crystallographic studies of both types of enzymes reveal that, other than containing a critical Mg^{2+} ion at the catalytic site, they are virtually unrelated to each other; indeed template-dependent nucleotide polymerizing enzymes seem to have arisen independently twice during the early evolution of cells. One lineage led to the modern DNA polymerases and reverse transcriptases discussed in Chapter 5, as well as to a few single-subunit RNA polymerases from viruses. The other lineage formed all of the modern cellular RNA polymerases (**Figure 6–10**), which we discuss in this chapter.

Cells Produce Several Types of RNA

The majority of genes carried in a cell's DNA specify the amino acid sequence of proteins; the RNA molecules that are copied from these genes (which ultimately direct the synthesis of proteins) are called **messenger RNA** (**mRNA**) molecules. The final product of a minority of genes, however, is the RNA itself. Careful analysis of the complete DNA sequence of the genome of the yeast *S. cerevisiae* has uncovered well over 750 genes (somewhat more than 10% of the total number of yeast genes) that produce RNA as their final product. These RNAs, like proteins, serve as enzymatic and structural components for a wide variety of processes in

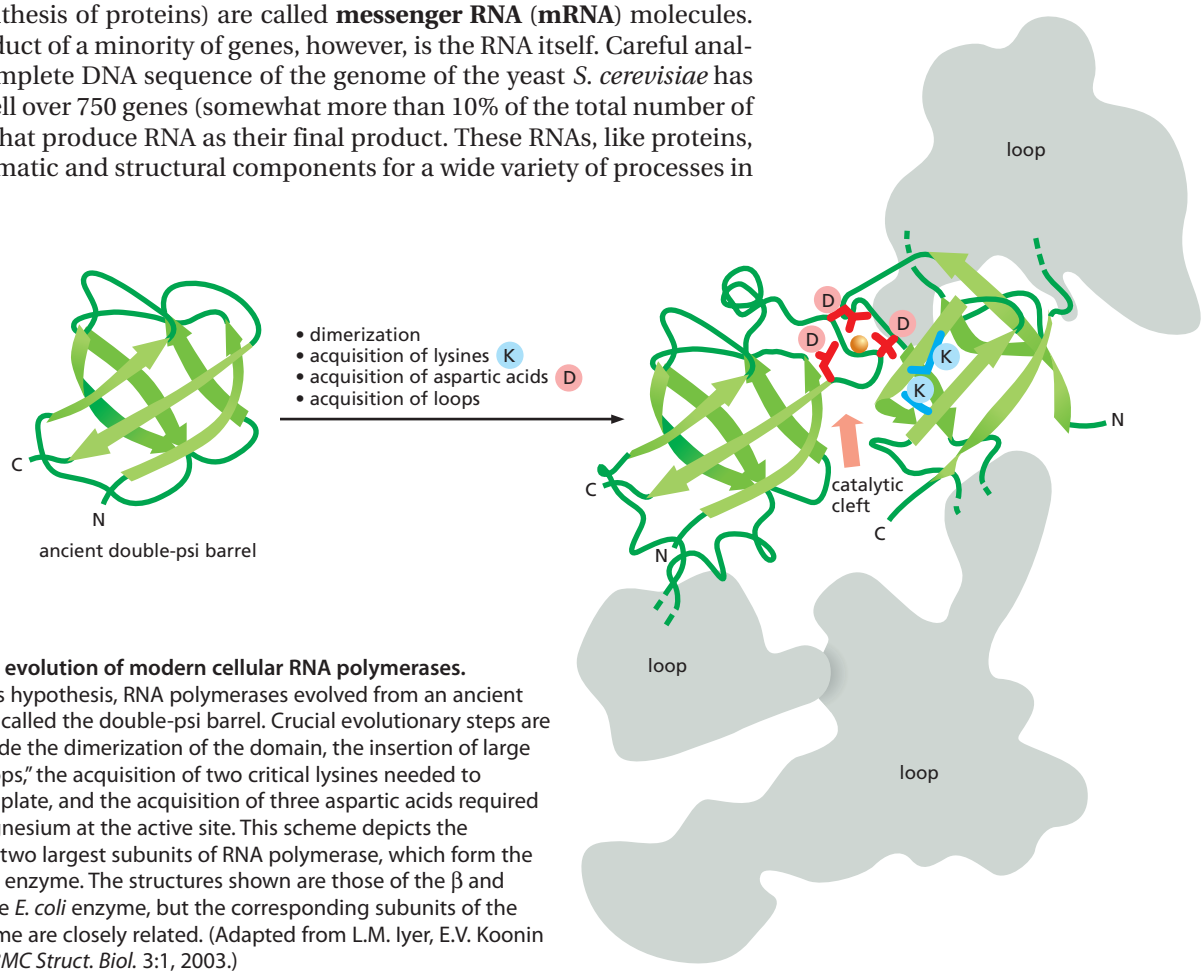


Figure 6–10 The evolution of modern cellular RNA polymerases. According to this hypothesis, RNA polymerases evolved from an ancient protein domain, called the double-psi barrel. Crucial evolutionary steps are thought to include the dimerization of the domain, the insertion of large polypeptide "loops," the acquisition of two critical lysines needed to position the template, and the acquisition of three aspartic acids required to chelate a magnesium at the active site. This scheme depicts the evolution of the two largest subunits of RNA polymerase, which form the active site of the enzyme. The structures shown are those of the β and β′ subunits of the *E. coli* enzyme, but the corresponding subunits of the eucaryotic enzyme are closely related. (Adapted from L.M. Iyer, E.V. Koonin and L. Aravind, *BMC Struct. Biol.* 3:1, 2003.)

the cell. In Chapter 5 we encountered one of those RNAs, the template carried by the enzyme telomerase. Although many of these noncoding RNAs are still mysterious, we shall see in this chapter that *small nuclear RNA (snRNA)* molecules direct the splicing of pre-mRNA to form mRNA, that *ribosomal RNA (rRNA)* molecules form the core of ribosomes, and that *transfer RNA (tRNA)* molecules form the adaptors that select amino acids and hold them in place on a ribosome for incorporation into protein. Finally, we shall see in Chapter 7 that *microRNA (miRNA)* molecules and *small interfering RNA (siRNA)* molecules serve as key regulators of eucaryotic gene expression (**Table 6–1**).

Each transcribed segment of DNA is called a *transcription unit*. In eucaryotes, a transcription unit typically carries the information of just one gene, and therefore codes for either a single RNA molecule or a single protein (or group of related proteins if the initial RNA transcript is spliced in more than one way to produce different mRNAs). In bacteria, a set of adjacent genes is often transcribed as a unit; the resulting mRNA molecule therefore carries the information for several distinct proteins.

Overall, RNA makes up a few percent of a cell's dry weight. Most of the RNA in cells is rRNA; mRNA comprises only 3–5% of the total RNA in a typical mammalian cell. The mRNA population is made up of tens of thousands of different species, and there are on average only 10–15 molecules of each species of mRNA present in each cell.

Signals Encoded in DNA Tell RNA Polymerase Where to Start and Stop

To transcribe a gene accurately, RNA polymerase must recognize where on the genome to start and where to finish. The way in which RNA polymerases perform these tasks differs somewhat between bacteria and eucaryotes. Because the processes in bacteria are simpler, we discuss them first.

The initiation of transcription is an especially important step in gene expression because it is the main point at which the cell regulates which proteins are to be produced and at what rate. The bacterial RNA polymerase core enzyme is a multisubunit complex that sythesizes RNA using a DNA template as a guide. A detachable subunit called *sigma (σ) factor* associates with the core enzyme and assists it in reading the signals in the DNA that tell it where to begin transcribing (**Figure 6–11**). Together, σ factor and core enzyme are known as the **RNA polymerase holoenzyme**; this complex adheres only weakly to bacterial DNA when

Table 6–1 Principal Types of RNAs Produced in Cells

TYPE OF RNA	FUNCTION
mRNAs	messenger RNAs, code for proteins
rRNAs	ribosomal RNAs, form the basic structure of the ribosome and catalyze protein synthesis
tRNAs	transfer RNAs, central to protein synthesis as adaptors between mRNA and amino acids
snRNAs	small nuclear RNAs, function in a variety of nuclear processes, including the splicing of pre-mRNA
snoRNAs	small nucleolar RNAs, used to process and chemically modify rRNAs
scaRNAs	small cajal RNAs, used to modify snoRNAs and snRNAs
miRNAs	microRNAs, regulate gene expression typically by blocking translation of selective mRNAs
siRNAs	small interfering RNAs, turn off gene expression by directing degradation of selective mRNAs and the establishment of compact chromatin structures
Other noncoding RNAs	function in diverse cell processes, including telomere synthesis, X-chromosome inactivation, and the transport of proteins into the ER

the two collide, and a holoenzyme typically slides rapidly along the long DNA molecule until it dissociates again. However, when the polymerase holoenzyme slides into a region on the DNA double helix called a **promoter**, a special sequence of nucleotides indicating the starting point for RNA synthesis, the polymerase binds tightly to this DNA. The polymerase holoenzyme, through its σ factor, recognizes the promoter DNA sequence by making specific contacts with the portions of the bases that are exposed on the outside of the helix (step 1 in Figure 6–11).

After the RNA polymerase holoenzyme binds tightly to the promoter DNA in this way, it opens up the double helix to expose a short stretch of nucleotides on each strand (step 2 in Figure 6–11). Unlike a DNA helicase reaction (see Figure 5–14), this limited opening of the helix does not require the energy of ATP hydrolysis. Instead, the polymerase and DNA both undergo reversible structural changes that result in a state more energetically favorable than that of the initial binding. With the DNA unwound, one of the two exposed DNA strands acts as a template for complementary base-pairing with incoming ribonucleotides, two of which are joined together by the polymerase to begin an RNA chain (step 3 in Figure 6–11). After the first ten or so nucleotides of RNA have been synthesized (a relatively inefficient process during which polymerase synthesizes and discards short RNA oligomers), the core enzyme breaks its interactions with the promoter DNA, weakens its interactions with σ factor, and begins to move down the DNA, synthesizing RNA (steps 4 and 5 in Figure 6–11). Chain elongation continues (at a speed of approximately 50 nucleotides/sec for bacterial RNA polymerases) until the enzyme encounters a second signal in the DNA, the **terminator** (described below), where the polymerase halts and releases both the newly made RNA chain and the DNA template (step 7 in Figure 6–11). After the polymerase core enzyme has been released at a terminator, it reassociates with a free σ factor to form a holoenzyme that can begin the process of transcription again.

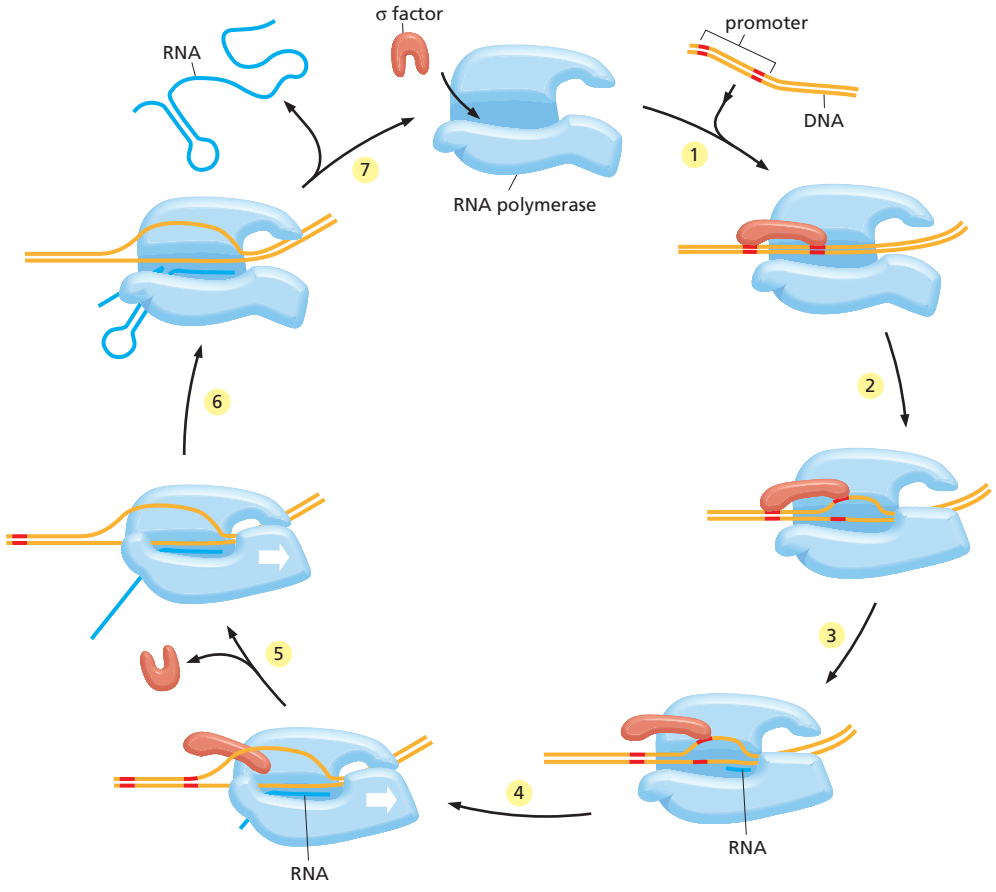
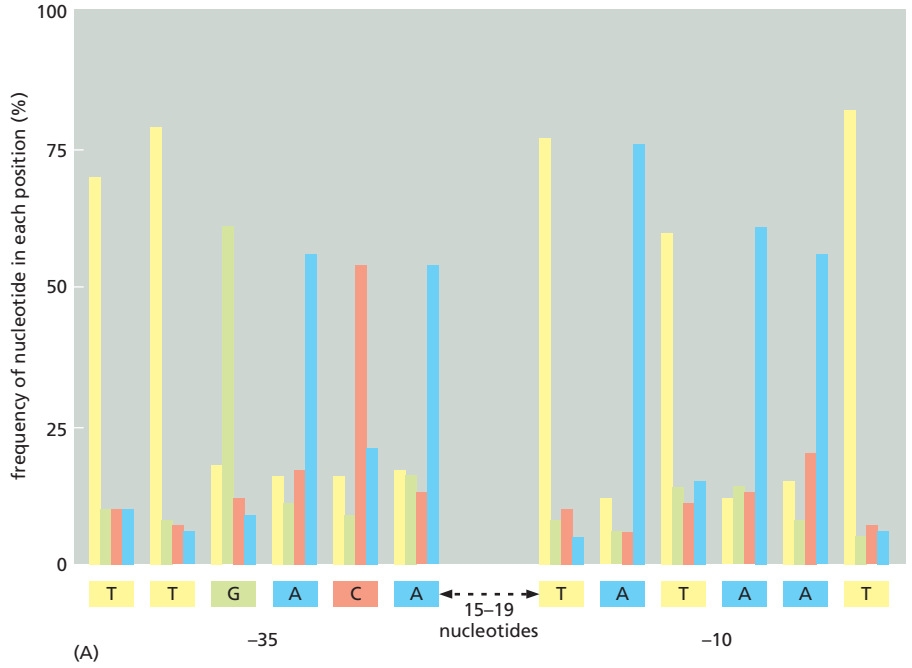


Figure 6–11 The transcription cycle of bacterial RNA polymerase. In step 1, the RNA polymerase holoenzyme (polymerase core enzyme plus σ factor) assembles and then locates a promoter (see Figure 6–12). The polymerase unwinds the DNA at the position at which transcription is to begin (step 2) and begins transcribing (step 3). This initial RNA synthesis (sometimes called "abortive initiation") is relatively inefficient. However, once RNA polymerase has managed to synthesize about 10 nucleotides of RNA, it breaks its interactions with the promoter DNA and weakens, and eventually breaks, its interaction with σ. The polymerase now shifts to the elongation mode of RNA synthesis (step 4), moving rightward along the DNA in this diagram. During the elongation mode (step 5), transcription is highly processive, with the polymerase leaving the DNA template and releasing the newly transcribed RNA only when it encounters a termination signal (steps 6 and 7). Termination signals are typically encoded in DNA, and many function by forming an RNA structure that destabilizes the polymerase's hold on the RNA (step 7). In bacteria, all RNA molecules are synthesized by a single type of RNA polymerase and the cycle depicted in the figure therefore applies to the production of mRNAs as well as structural and catalytic RNAs. (Adapted from a figure courtesy of Robert Landick.)



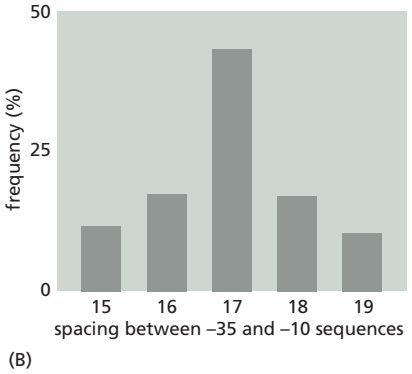


Figure 6–12 Consensus sequence for the major class of *E. coli* promoters. (A) The promoters are characterized by two hexameric DNA sequences, the –35 sequence and the –10 sequence named for their approximate location relative to the start point of transcription (designated +1). For convenience, the nucleotide sequence of a single strand of DNA is shown; in reality the RNA polymerase recognizes the promoter as double-stranded DNA. On the basis of a comparison of 300 promoters, the frequencies of the four nucleotides at each position in the –35 and –10 hexamers are given. The consensus sequence, shown *below* the graph, reflects the most common nucleotide found at each position in the collection of promoters. The sequence of nucleotides between the –35 and –10 hexamers shows no significant similarities among promoters. (B) The distribution of spacing between the –35 and –10 hexamers found in *E. coli* promoters.

The information displayed in these two graphs applies to *E. coli* promoters that are recognized by RNA polymerase and the major σ factor (designated σ^{70}). As we shall see in the next chapter, bacteria also contain minor σ factors, each of which recognizes a different promoter sequence. Some particularly strong promoters recognized by RNA polymerase and σ^{70} have an additional sequence, located upstream (to the *left,* in the figure) of the –35 hexamer, which is recognized by another subunit of RNA polymerase.

The process of transcription initiation is complex and requires that the RNA polymerase holoenzyme and the DNA undergo a series of conformational changes. We can view these changes as opening up and positioning the DNA in the active site followed by a successive tightening of the enzyme around the DNA and RNA to ensure that it does not dissociate before it has finished transcribing a gene. If an RNA polymerase does dissociate prematurely, it cannot resume synthesis but must start over again at the promoter.

How do the termination signals in the DNA stop the elongating polymerase? For most bacterial genes a termination signal consists of a string of A–T nucleotide pairs preceded by a two-fold symmetric DNA sequence, which, when transcribed into RNA, folds into a "hairpin" structure through Watson–Crick base-pairing (see Figure 6–11). As the polymerase transcribes across a terminator, the formation of the hairpin may help to "pull" the RNA transcript from the active site. The DNA–RNA hybrid in the active site, which is held together at terminators predominantly by U–A base pairs (which are less stable than G–C base pairs because they form two rather than three hydrogen bonds per base pair), is not strong enough to hold the RNA in place, and it dissociates causing the release of the polymerase from the DNA (step 7 in Figure 6–11). Thus, in some respects, transcription termination seems to involve a reversal of the structural transitions that happen during initiation. The process of termination also is an example of a common theme in this chapter: the folding of RNA into specific structures affects many steps in decoding the genome.

Transcription Start and Stop Signals Are Heterogeneous in Nucleotide Sequence

As we have just seen, the processes of transcription initiation and termination involve a complicated series of structural transitions in protein, DNA, and RNA molecules. The signals encoded in DNA that specify these transitions are often difficult for researchers to recognize. Indeed, a comparison of many different bacterial promoters reveals a surprising degree of variation. Nevertheless, they all contain related sequences, reflecting in part aspects of the DNA that are recognized directly by the σ factor. These common features are often summarized in the form of a *consensus sequence* (**Figure 6–12**). A **consensus nucleotide sequence** is derived by comparing many sequences with the same basic function and tallying up the most common nucleotide found at each position. It

therefore serves as a summary or "average" of a large number of individual nucleotide sequences.

The DNA sequences of individual bacterial promoters differ in ways that determine their strength (the number of initiation events per unit time of the promoter). Evolutionary processes have fine-tuned each to initiate as often as necessary and have thereby created a wide spectrum of promoters. Promoters for genes that code for abundant proteins are much stronger than those associated with genes that encode rare proteins, and their nucleotide sequences are responsible for these differences.

Like bacterial promoters, transcription terminators also have a wide range of sequences, with the potential to form a simple hairpin RNA structure being the most important common feature. Since an almost unlimited number of nucleotide sequences have this potential, terminator sequences are even more heterogeneous than promoter sequences.

We have discussed bacterial promoters and terminators in some detail to illustrate an important point regarding the analysis of genome sequences. Although we know a great deal about bacterial promoters and terminators and can construct consensus sequences that summarize their most salient features, their variation in nucleotide sequence makes it difficult to definitively locate them simply by analysis of the nucleotide sequence of a genome. It is even more difficult to locate analogous sequences in eucaryotic genomes, due in part to the excess DNA carried in them. Often, we need additional information, some of it from direct experimentation, to locate and accurately interpret the short DNA signals contained in genomes.

Since DNA is double-stranded, two different RNA molecules could in principle be transcribed from any gene, using each of the two DNA strands as a template. However, a gene typically has only a single promoter, and because the promoter's nucleotide sequence is asymmetric (see Figure 6–12), the polymerase can bind in only one orientation. The polymerase synthesizes RNA in the 5′-to-3′ direction, and it can therefore only transcribe one strand per gene (**Figure 6–13**). Genome sequences reveal that the DNA strand used as the template for RNA synthesis varies from gene to gene depending on the location and orientation of the promoter (**Figure 6–14**).

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

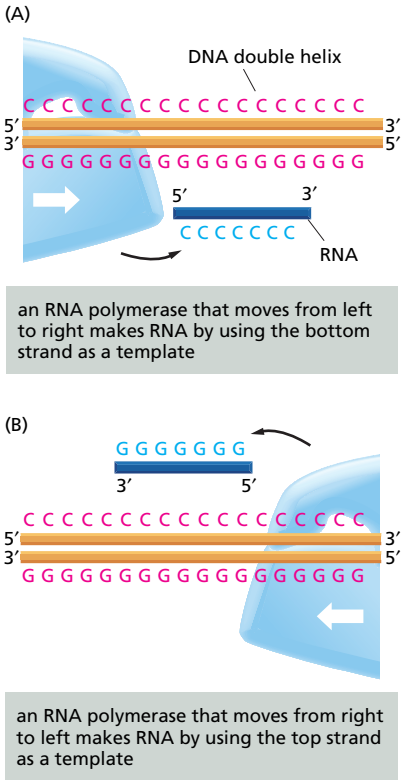


Figure 6–13 The importance of RNA polymerase orientation. The DNA strand serving as template must be traversed in a 3′-to-5′ direction. Thus, the direction of RNA polymerase movement determines which of the two DNA strands is to serve as a template for the synthesis of RNA, as shown in (A) and (B). Polymerase direction is, in turn, determined by the orientation of the promoter sequence, the site at which the RNA polymerase begins transcription.

Transcription Initiation in Eucaryotes Requires Many Proteins

In contrast to bacteria, which contain a single type of RNA polymerase, eucaryotic nuclei have three: *RNA polymerase I, RNA polymerase II,* and *RNA polymerase III.* The three polymerases are structurally similar to one another (and to the bacterial enzyme) and share some common subunits, but they transcribe different types of genes (**Table 6–2**). RNA polymerases I and III transcribe the genes encoding transfer RNA, ribosomal RNA, and various small RNAs. RNA polymerase II transcribes most genes, including all those that encode proteins, and our subsequent discussion therefore focuses on this enzyme.

Although eucaryotic RNA polymerase II has many structural similarities to bacterial RNA polymerase (**Figure 6–15**), there are several important differences in the way in which the bacterial and eucaryotic enzymes function, two of which concern us immediately.

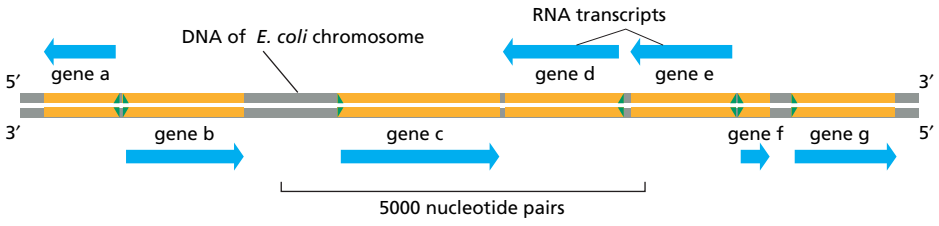


Figure 6–14 Directions of transcription along a short portion of a bacterial chromosome. Some genes are transcribed using one DNA strand as a template, while others are transcribed using the other DNA strand. The direction of transcription is determined by the promoter at the beginning of each gene *(green arrowheads).* This diagram shows approximately 0.2% (9000 base pairs) of the *E. coli* chromosome. The genes transcribed from *left* to *right* use the bottom DNA strand as the template; those transcribed from *right* to *left* use the top strand as the template.

Table 6–2 The Three RNA Polymerases in Eucaryotic Cells

TYPE OF POLYMERASE	GENES TRANSCRIBED
RNA polymerase I	5.8S, 18S, and 28S rRNA genes
RNA polymerase II	all protein-coding genes, plus snoRNA genes, miRNA genes, siRNA genes, and most snRNA genes
RNA polymerase III	tRNA genes, 5S rRNA genes, some snRNA genes and genes for other small RNAs

The rRNAs are named according to their "S" values, which refer to their rate of sedimentation in an ultracentrifuge. The larger the S value, the larger the rRNA.

1. While bacterial RNA polymerase requires only a single additional protein (σ factor) for transcription initiation to occur *in vitro*, eucaryotic RNA polymerases require many additional proteins, collectively called the *general transcription factors*.
2. Eucaryotic transcription initiation must deal with the packing of DNA into nucleosomes and higher-order forms of chromatin structure, features absent from bacterial chromosomes.

RNA Polymerase II Requires General Transcription Factors

The **general transcription factors** help to position eucaryotic RNA polymerase correctly at the promoter, aid in pulling apart the two strands of DNA to allow transcription to begin, and release RNA polymerase from the promoter into the elongation mode once transcription has begun. <CTAT> The proteins are "general" because they are needed at nearly all promoters used by RNA polymerase II; consisting of a set of interacting proteins, they are designated as *TFII* (for transcription factor for polymerase II), and are denoted arbitrarily as TFIIB, TFIID, and so on. In a broad sense, the eucaryotic general transcription factors carry out functions equivalent to those of the σ factor in bacteria; indeed, portions of TFIIF have the same three-dimensional structure as the equivalent portions of σ.

Figure 6–16 illustrates how the general transcription factors assemble at promoters used by RNA polymerase II, and **Table 6–3** summarizes their activities. The assembly process begins when the general transcription factor TFIID binds to a short double-helical DNA sequence primarily composed of T and A nucleotides. For this reason, this sequence is known as the TATA sequence, or **TATA box**, and the subunit of TFIID that recognizes it is called TBP (for TATA-binding protein). The TATA box is typically located 25 nucleotides upstream from the transcription start site. It is not the only DNA sequence that signals the start of transcription (**Figure 6–17**), but for most polymerase II promoters it is the most important. The binding of TFIID causes a large distortion in the DNA

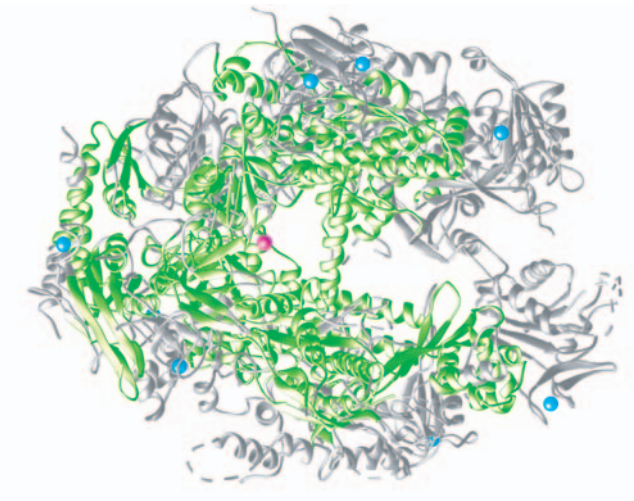


Figure 6–15 Structural similarity between a bacterial RNA polymerase and a eucaryotic RNA polymerase II. Regions of the two RNA polymerases that have similar structures are indicated in *green*. The eucaryotic polymerase is larger than the bacterial enzyme (12 subunits instead of 5), and some of the additional regions are shown in *gray*. The *blue* spheres represent Zn atoms that serve as structural components of the polymerases, and the *red* sphere represents the Mg atom present at the active site, where polymerization takes place. The RNA polymerases in all modern-day cells (bacteria, archaea, and eucaryotes) are closely related, indicating that the basic features of the enzyme were in place before the divergence of the three major branches of life. (Courtesy of P. Cramer and R. Kornberg.)

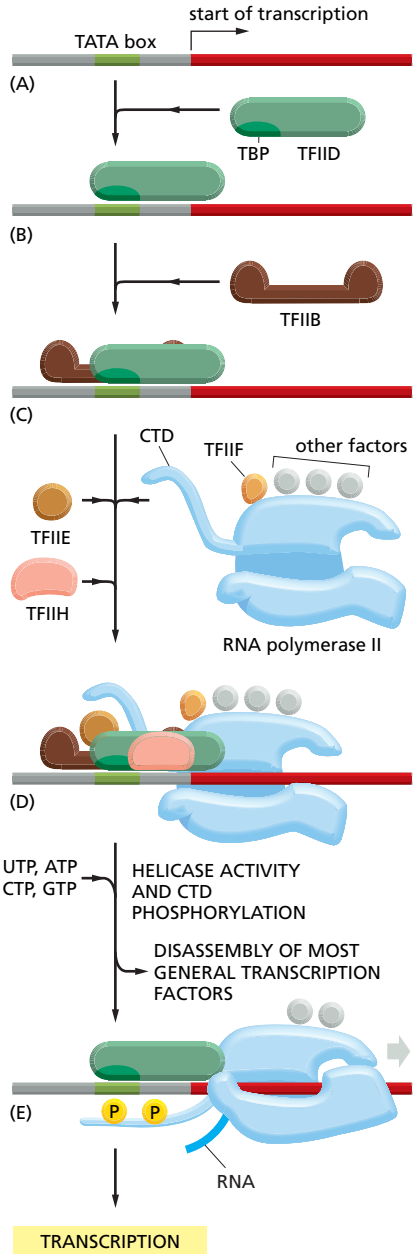


Figure 6–16 Initiation of transcription of a eucaryotic gene by RNA polymerase II. To begin transcription, RNA polymerase requires several general transcription factors. (A) The promoter contains a DNA sequence called the TATA box, which is located 25 nucleotides away from the site at which transcription is initiated. (B) Through its subunit TBP, TFIID recognizes and binds the TATA box, which then enables the adjacent binding of TFIIB (C). For simplicity the DNA distortion produced by the binding of TFIID (see Figure 6–18) is not shown. (D) The rest of the general transcription factors, as well as the RNA polymerase itself, assemble at the promoter. (E) TFIIH then uses ATP to pry apart the DNA double helix at the transcription start point, locally exposing the template strand. TFIIH also phosphorylates RNA polymerase II, changing its conformation so that the polymerase is released from the general factors and can begin the elongation phase of transcription. As shown, the site of phosphorylation is a long C-terminal polypeptide tail, also called the C-terminal domain (CTD), that extends from the polymerase molecule. The assembly scheme shown in the figure was deduced from experiments performed *in vitro*, and the exact order in which the general transcription factors assemble on promoters may vary from gene to gene *in vivo*. The general transcription factors have been highly conserved in evolution; some of those from human cells can be replaced in biochemical experiments by the corresponding factors from simple yeasts.

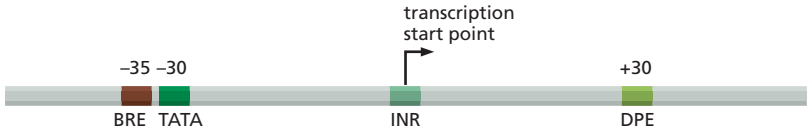
of the TATA box (**Figure 6–18**). This distortion is thought to serve as a physical landmark for the location of an active promoter in the midst of a very large genome, and it brings DNA sequences on both sides of the distortion together to allow for subsequent protein assembly steps. Other factors then assemble, along with RNA polymerase II, to form a complete *transcription initiation complex* (see Figure 6–16). The most complicated of the general transcription factors is TFIIH. Consisting of 9 subunits, it is nearly as large as RNA polymerase II itself and, as we shall see shortly, performs several enzymatic steps needed for the initiation of transcription.

After forming a transcription initiation complex on the promoter DNA, RNA polymerase II must gain access to the template strand at the transcription start point. TFIIH, which contains a DNA helicase as one of its subunits, makes this step possible by hydrolyzing ATP and unwinding the DNA, thereby exposing the template strand. Next, RNA polymerase II, like the bacterial polymerase, remains at the promoter synthesizing short lengths of RNA until it undergoes a series of conformational changes that allow it to move away from the promoter and enter the elongation phase of transcription. A key step in this transition is the addition of phosphate groups to the "tail" of the RNA polymerase (known as the CTD or C-terminal domain). In humans, the CTD consists of 52 tandem repeats of a seven-amino-acid sequence, which extend from the RNA polymerase core structure. During transcription initiation, the serine located at the

Table 6–3 The General Transcription Factors Needed for Transcription Initiation by Eucaryotic RNA Polymerase II

NAME	NUMBER OF SUBUNITS	ROLES IN TRANSITION INITIATION
TFIID		
TBP subunit	1	recognizes TATA box
TAF subunits	~11	recognizes other DNA sequences near the transcription start point; regulates DNA-binding by TBP
TFIIB	1	recognizes BRE element in promoters; accurately positions RNA polymerase at the start site of transcription
TFIIF	3	stabilizes RNA polymerase interaction with TBP and TFIIB; helps attract TFIIE and TFIIH
TFIIE	2	attracts and regulates TFIIH
TFIIH	9	unwinds DNA at the transcription start point, phosphorylates Ser5 of the RNA polymerase CTD; releases RNA polymerase from the promoter

TFIID is composed of TBP and ~11 additional subunits called TAFs (TBP-associated factors); CTD, C-terminal domain.



element	consensus sequence	general transcription factor
BRE	G/C G/C G/A C G C C	TFIIB
TATA	T A T A A/T A A/T	TBP
INR	C/T C/T A N T/A C/T C/T	TFIID
DPE	A/G G A/T C G T G	TFIID

Figure 6–17 Consensus sequences found in the vicinity of eucaryotic RNA polymerase II start points. The name given to each consensus sequence *(first column)* and the general transcription factor that recognizes it *(last column)* are indicated. N indicates any nucleotide, and two nucleotides separated by a slash indicate an equal probability of either nucleotide at the indicated position. In reality, each consensus sequence is a shorthand representation of a histogram similar to that of Figure 6–12.

For most RNA polymerase II transcription start points, only two or three of the four sequences are present. For example, many polymerase II promoters have a TATA box sequence, but those that do not typically have a "strong" INR sequence. Although most of the DNA sequences that influence transcription initiation are located upstream of the transcription start point, a few, such as the DPE shown in the figure, are located in the transcribed region.

fifth position in the repeat sequence (Ser5) is phosphorylated by TFIIH, which contains a protein kinase in another of its subunits (see Figure 6–16D and E). The polymerase can then disengage from the cluster of general transcription factors. During this process, it undergoes a series of conformational changes that tighten its interaction with DNA, and it acquires new proteins that allow it to transcribe for long distances, and in some cases for many hours, without dissociating from DNA.

Once the polymerase II has begun elongating the RNA transcript, most of the general transcription factors are released from the DNA so that they are available to initiate another round of transcription with a new RNA polymerase molecule. As we see shortly, the phosphorylation of the tail of RNA polymerase II also causes components of the RNA-processing machinery to load onto the polymerase and thus be positioned to modify the newly transcribed RNA as it emerges from the polymerase.

Polymerase II Also Requires Activator, Mediator, and Chromatin-Modifying Proteins

Studies of the behavior of RNA polymerase II and its general transcription factors on purified DNA templates *in vitro* established the model for transcription initiation just described. However, as discussed in Chapter 4, DNA in eucaryotic cells is packaged into nucleosomes, which are further arranged in higher-order

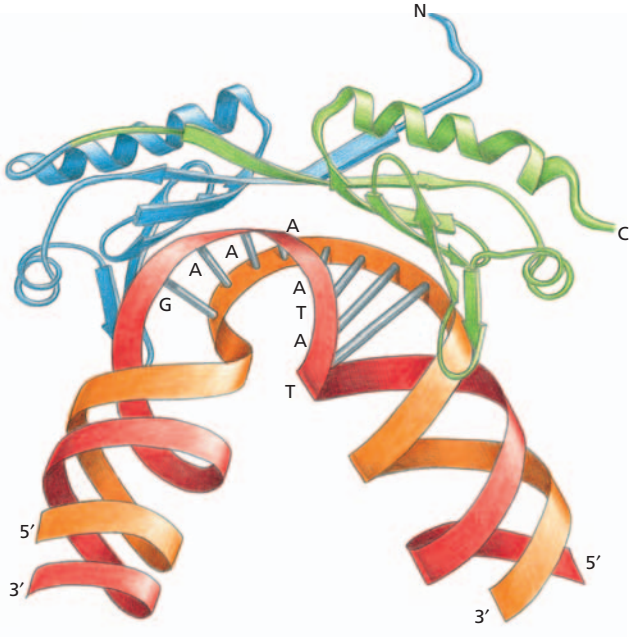


Figure 6–18 Three-dimensional structure of TBP (TATA-binding protein) bound to DNA. The TBP is the subunit of the general transcription factor TFIID that is responsible for recognizing and binding to the TATA box sequence in the DNA *(red)*. The unique DNA bending caused by TBP—two kinks in the double helix separated by partly unwound DNA—may serve as a landmark that helps to attract the other general transcription factors. TBP is a single polypeptide chain that is folded into two very similar domains *(blue* and *green)*. (Adapted from J.L. Kim et al., *Nature* 365:520–527, 1993. With permission from Macmillan Publishers Ltd.)

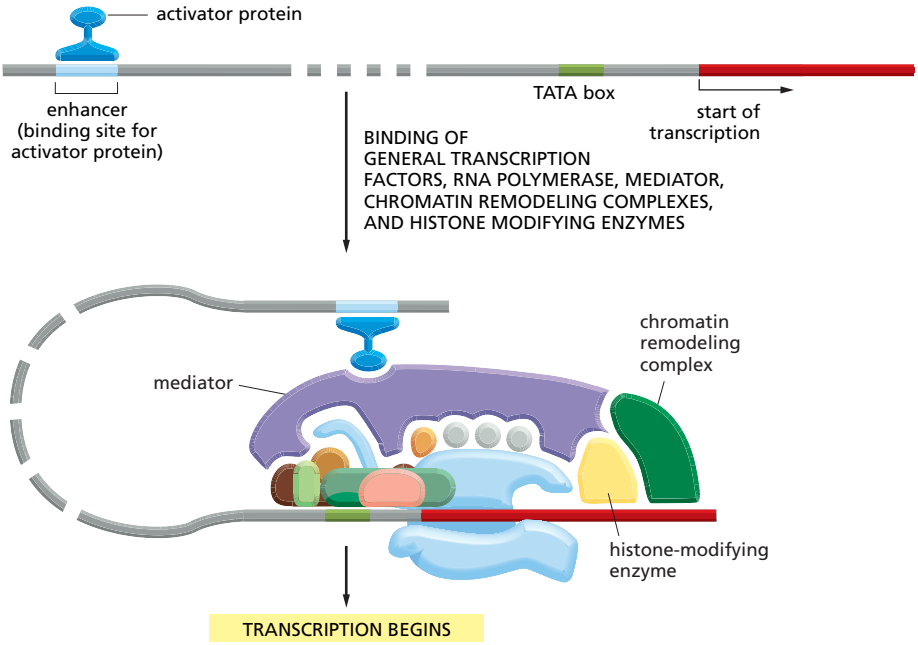


Figure 6–19 Transcription initiation by RNA polymerase II in a eucaryotic cell. Transcription initiation *in vivo* requires the presence of transcriptional activator proteins. As described in Chapter 7, these proteins bind to specific short sequences in DNA. Although only one is shown here, a typical eucaryotic gene has many activator proteins, which together determine its rate and pattern of transcription. Sometimes acting from a distance of several thousand nucleotide pairs (indicated by the dashed DNA molecule), these gene regulatory proteins help RNA polymerase, the general transcription factors, and the mediator all to assemble at the promoter. In addition, activators attract ATP-dependent chromatin remodeling complexes and histone acetylases.

As discussed in Chapter 4, the "default" state of chromatin is probably the 30-nm filament (see Figure 4–22), and this is likely to be a form of DNA upon which transcription is initiated. For simplicity, it is not shown in the figure.

chromatin structures. As a result, transcription initiation in a eucaryotic cell is more complex and requires even more proteins than it does on purified DNA. First, gene regulatory proteins known as *transcriptional activators* must bind to specific sequences in DNA and help to attract RNA polymerase II to the start point of transcription (**Figure 6–19**). We discuss the role of activators in Chapter 7, because they are one of the main ways in which cells regulate expression of their genes. Here we simply note that their presence on DNA is required for transcription initiation in a eucaryotic cell. Second, eucaryotic transcription initiation *in vivo* requires the presence of a protein complex known as *Mediator*, which allows the activator proteins to communicate properly with the polymerase II and with the general transcription factors. Finally, transcription initiation in a eucaryotic cell typically requires the local recruitment of chromatin-modifying enzymes, including chromatin remodeling complexes and histone-modifying enzymes. As discussed in Chapter 4, both types of enzymes can allow greater access to the DNA present in chromatin, and by doing so, they facilitate the assembly of the transcription initiation machinery onto DNA. We will revisit the role of these enzymes in transcription initiation in Chapter 7.

As illustrated in Figure 6–19, many proteins (well over 100 individual subunits) must assemble at the start point of transcription to initiate transcription in a eucaryotic cell. The order of assembly of these proteins does not seem to follow a prescribed pathway; rather, the order differs from gene to gene. Indeed, some of these different protein complexes may interact with each other away from the DNA and be brought to DNA as preformed subassemblies. To begin transcribing, RNA polymerase II must be released from this large complex of proteins, and, in addition to the steps described in Figure 6–16, this often requires the *in situ* proteolysis of the activator protein. We return to some of these issues in Chapter 7, where we discuss how eucaryotic cells can regulate the process of transcription initiation.

Transcription Elongation Produces Superhelical Tension in DNA

Once it has initiated transcription, RNA polymerase does not proceed smoothly along a DNA molecule; rather, it moves jerkily, pausing at some sequences and rapidly transcribing through others. Elongating RNA polymerases, both bacterial and eucaryotic, are associated with a series of *elongation factors*, proteins that decrease the likelihood that RNA polymerase will dissociate before it reaches the end of a gene. These factors typically associate with RNA polymerase

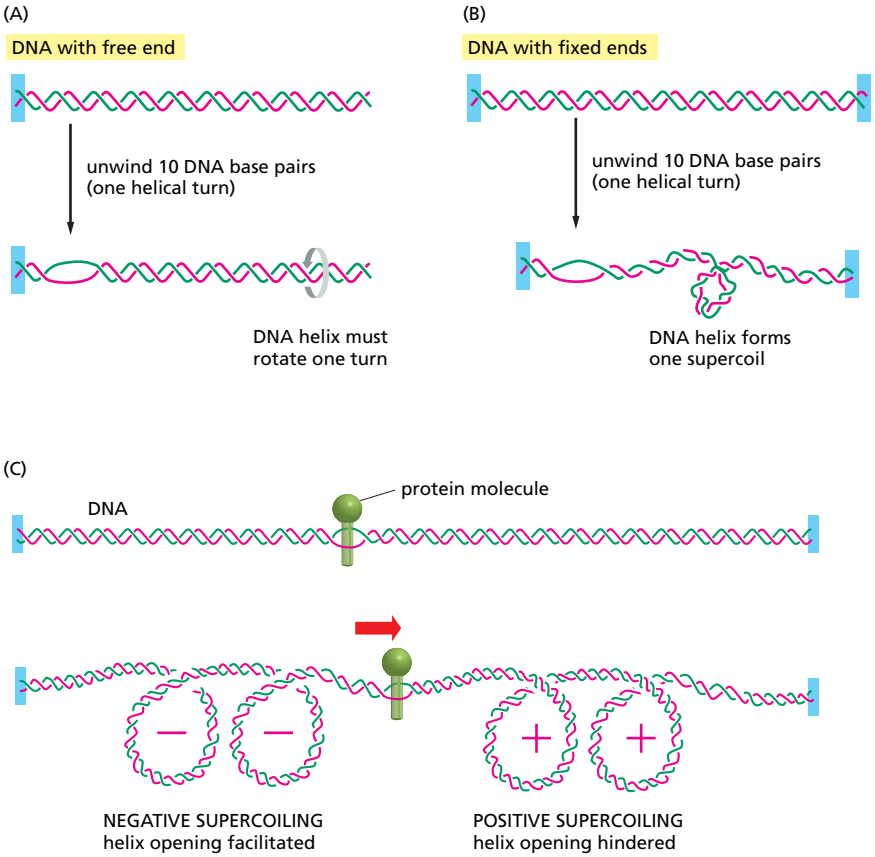


Figure 6–20 Superhelical tension in DNA causes DNA supercoiling. (A) For a DNA molecule with one free end (or a nick in one strand that serves as a swivel), the DNA double helix rotates by one turn for every 10 nucleotide pairs opened. (B) If rotation is prevented, superhelical tension is introduced into the DNA by helix opening. One way of accommodating this tension would be to increase the helical twist from 10 to 11 nucleotide pairs per turn in the double helix that remains; the DNA helix, however, resists such a deformation in a springlike fashion, preferring to relieve the superhelical tension by bending into supercoiled loops. As a result, one DNA supercoil forms in the DNA double helix for every 10 nucleotide pairs opened. The supercoil formed in this case is a positive supercoil. (C) Supercoiling of DNA is induced by a protein tracking through the DNA double helix. The two ends of the DNA shown here are unable to rotate freely relative to each other, and the protein molecule is assumed also to be prevented from rotating freely as it moves. Under these conditions, the movement of the protein causes an excess of helical turns to accumulate in the DNA helix ahead of the protein and a deficit of helical turns to arise in the DNA behind the protein, as shown.

shortly after initiation and help polymerases to move through the wide variety of different DNA sequences that are found in genes. Eucaryotic RNA polymerases must also contend with chromatin structure as they move along a DNA template, and they are typically aided by ATP-dependent chromatin remodeling complexes (see pp. 215–216). These complexes may move with the polymerase or may simply seek out and rescue the occasional stalled polymerase. In addition, some elongation factors associated with eucaryotic RNA polymerase facilitate transcription through nucleosomes without requiring additional energy. It is not yet understood in detail how this is accomplished, but these proteins can transiently dislodge H2A–H2B dimers from the nucleosome core, replacing them as the polymerase moves through the nucleosome.

There is yet another barrier to elongating polymerases, both bacterial and eucaryotic. To discuss this issue, we need first to consider a subtle property inherent in the DNA double helix called **DNA supercoiling**. DNA supercoiling represents a conformation that DNA adopts in response to superhelical tension; conversely, creating various loops or coils in the helix can create such tension. **Figure 6–20** illustrates the topological constraints that cause DNA supercoiling. There are approximately 10 nucleotide pairs for every helical turn in a DNA double helix. Imagine a helix whose two ends are fixed with respect to each other (as they are in a DNA circle, such as a bacterial chromosome, or in a tightly clamped loop, as is thought to exist in eucaryotic chromosomes). In this case, one large DNA supercoil will form to compensate for each 10 nucleotide pairs that are opened (unwound). The formation of this supercoil is energetically favorable because it restores a normal helical twist to the base-paired regions that remain, which would otherwise need to be overwound because of the fixed ends.

RNA polymerase also creates superhelical tension as it moves along a stretch of DNA that is anchored at its ends (see Figure 6–20C). As long as the polymerase is not free to rotate rapidly (and such rotation is unlikely given the size of RNA polymerases and their attached transcripts), a moving polymerase generates

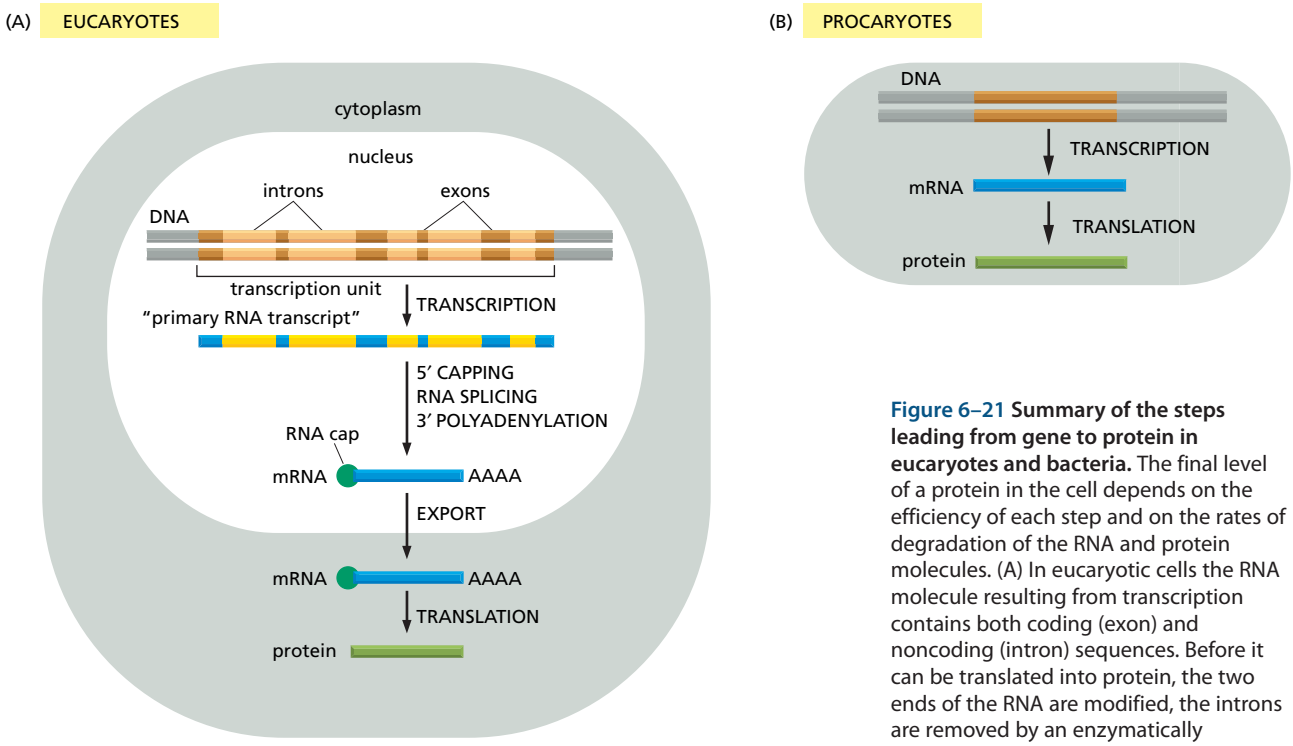


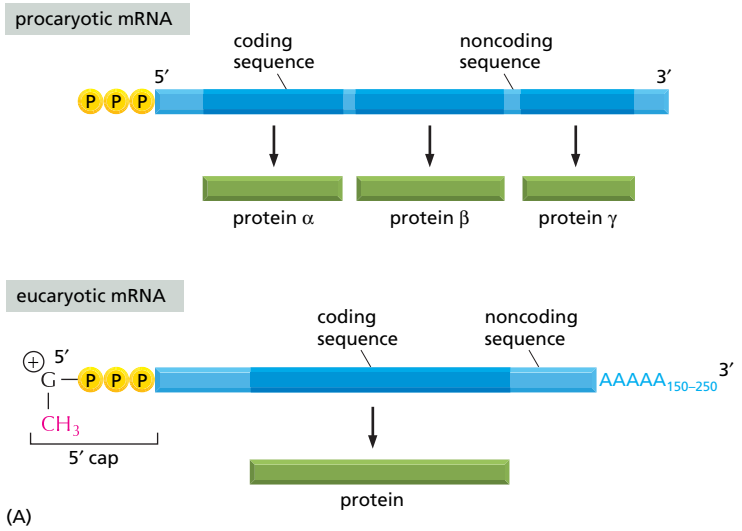
Figure 6–21 Summary of the steps leading from gene to protein in eucaryotes and bacteria. The final level of a protein in the cell depends on the efficiency of each step and on the rates of degradation of the RNA and protein molecules. (A) In eucaryotic cells the RNA molecule resulting from transcription contains both coding (exon) and noncoding (intron) sequences. Before it can be translated into protein, the two ends of the RNA are modified, the introns are removed by an enzymatically catalyzed RNA splicing reaction, and the resulting mRNA is transported from the nucleus to the cytoplasm. Although the steps in this figure are depicted as occurring one at a time, in a sequence, in reality they can occur concurrently. For example, the RNA cap is added and splicing typically begins before transcription has been completed. Because of the coupling between transcription and RNA processing, primary transcripts—the RNAs that would, in theory, be produced if no processing had occurred—are found only rarely. (B) In procaryotes the production of mRNA is much simpler. The 5′ end of an mRNA molecule is produced by the initiation of transcription, and the 3′ end is produced by the termination of transcription. Since procaryotic cells lack a nucleus, transcription and translation take place in a common compartment. In fact, the translation of a bacterial mRNA often begins before its synthesis has been completed.

positive superhelical tension in the DNA in front of it and negative helical tension behind it. For eucaryotes, this situation is thought to provide a bonus: the positive superhelical tension ahead of the polymerase makes the DNA helix more difficult to open, but this tension should facilitate the unwrapping of DNA in nucleosomes, as the release of DNA from the histone core helps to relax positive superhelical tension.

Any protein that propels itself alone along a DNA strand of a double helix tends to generate superhelical tension. In eucaryotes, DNA topoisomerase enzymes rapidly remove this superhelical tension (see p. 278). But in bacteria a specialized topoisomerase called *DNA gyrase* uses the energy of ATP hydrolysis to pump supercoils continuously into the DNA, thereby maintaining the DNA under constant tension. These are *negative supercoils,* having the opposite handedness from the *positive supercoils* that form when a region of DNA helix opens (see Figure 6–20B). Whenever a region of helix opens, it removes these negative supercoils from bacterial DNA, reducing the superhelical tension. DNA gyrase therefore makes the opening of the DNA helix in bacteria energetically favorable compared with helix opening in DNA that is not supercoiled. For this reason, it usually facilitates those genetic processes in bacteria, including the initiation of transcription by bacterial RNA polymerase, that require helix opening (see Figure 6–11).

Transcription Elongation in Eucaryotes Is Tightly Coupled to RNA Processing

We have seen that bacterial mRNAs are synthesized solely by the RNA polymerase starting and stopping at specific spots on the genome. The situation in eucaryotes is substantially different. In particular, transcription is only the first of several steps needed to produce an mRNA. Other critical steps are the covalent modification of the ends of the RNA and the removal of *intron sequences* that are discarded from the middle of the RNA transcript by the process of *RNA splicing* (**Figure 6–21**).



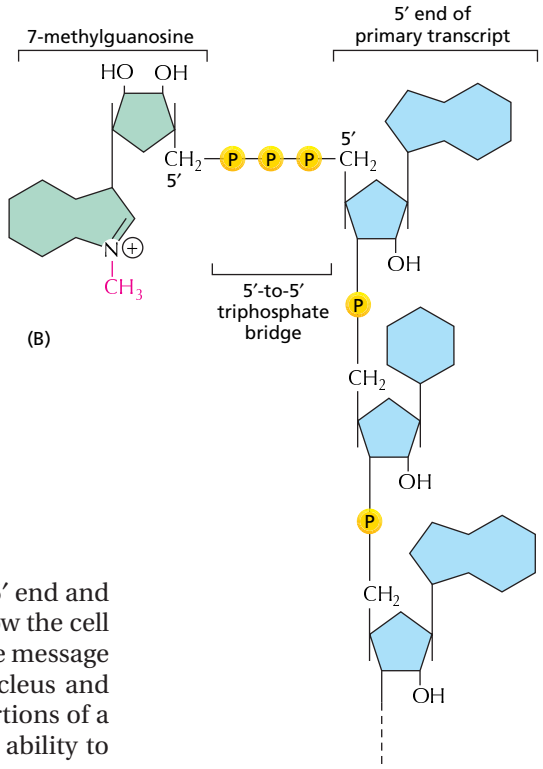


Figure 6–22 A comparison of the structures of procaryotic and eucaryotic mRNA molecules. (A) The 5′ and 3′ ends of a bacterial mRNA are the unmodified ends of the chain synthesized by the RNA polymerase, which initiates and terminates transcription at those points, respectively. The corresponding ends of a eucaryotic mRNA are formed by adding a 5′ cap and by cleavage of the pre-mRNA transcript and the addition of a poly-A tail, respectively. The figure also illustrates another difference between the procaryotic and eucaryotic mRNAs: bacterial mRNAs can contain the instructions for several different proteins, whereas eucaryotic mRNAs nearly always contain the information for only a single protein. (B) The structure of the cap at the 5′ end of eucaryotic mRNA molecules. Note the unusual 5′-to-5′ linkage of the 7-methyl G to the remainder of the RNA. Many eucaryotic mRNAs carry an additional modification: the 2′-hydroxyl group on the second ribose sugar in the mRNA is methylated (not shown).

Both ends of eucaryotic mRNAs are modified: by *capping* on the 5′ end and by *polyadenylation* of the 3′ end (**Figure 6–22**). These special ends allow the cell to assess whether both ends of an mRNA molecule are present (and the message is therefore intact) before it exports the RNA sequence from the nucleus and translates it into protein. RNA splicing joins together the different portions of a protein coding sequence, and it provides higher eucaryotes with the ability to synthesize several different proteins from the same gene.

An ingenious mechanism couples all of the above RNA processing steps to transcription elongation. As discussed previously, a key step in transcription initiation by RNA polymerase II is the phosphorylation of the RNA polymerase II tail, called the CTD (C-terminal domain). This phosphorylation proceeds gradually as the RNA polymerase initiates transcription and moves along the DNA. It not only helps dissociate the RNA polymerase II from other proteins present at the start point of transcription, but also allows a new set of proteins to associate with the RNA polymerase tail that function in transcription elongation and RNA processing. As discussed next, some of these processing proteins seem to "hop" from the polymerase tail onto the nascent RNA molecule to begin processing it as it emerges from the RNA polymerase. Thus, we can view RNA polymerase II in its elongation mode as an RNA factory that both transcribes DNA into RNA and processes the RNA it produces (**Figure 6–23**). Fully extended, the CTD is nearly 10 times longer than the remainder of RNA polymerase and, in effect, it serves as a tether, holding a variety of proteins close by until they are needed. This strategy, which speeds up the rate of subsequent reactions, is one commonly observed in the cell (see Figures 4–69 and 16–38).

RNA Capping Is the First Modification of Eucaryotic Pre-mRNAs

As soon as RNA polymerase II has produced about 25 nucleotides of RNA, the 5′ end of the new RNA molecule is modified by addition of a cap that consists of a modified guanine nucleotide (see Figure 6–22B). Three enzymes, acting in succession, perform the capping reaction: one (a phosphatase) removes a phosphate from the 5′ end of the nascent RNA, another (a guanyl transferase) adds a GMP in a reverse linkage (5′ to 5′ instead of 5′ to 3′), and a third (a methyl transferase) adds a methyl group to the guanosine (**Figure 6–24**). Because all three enzymes bind to the RNA polymerase tail phosphorylated at serine-5 position, the modification added by TFIIH during transcription initiation, they are poised to modify the 5′ end of the nascent transcript as soon as it emerges from the polymerase.

The 5′-methyl cap signifies the 5′ end of eucaryotic mRNAs, and this landmark helps the cell to distinguish mRNAs from the other types of RNA molecules present in the cell. For example, RNA polymerases I and III produce uncapped

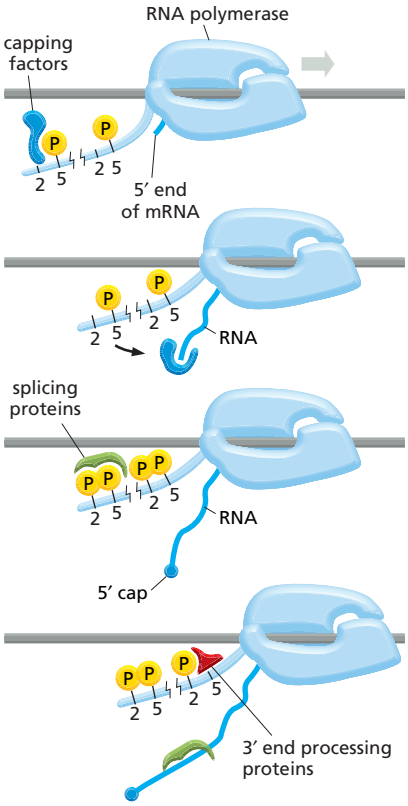


Figure 6–23 Eucaryotic RNA polymerase II as an "RNA factory." As the polymerase transcribes DNA into RNA, it carries pre-mRNA-processing proteins on its tail that are transferred to the nascent RNA at the appropriate time. The tail, known as the CTD, contains 52 tandem repeats of a seven amino acid sequence, and there are two serines in each repeat. The capping proteins first bind to the RNA polymerase tail when it is phosphorylated on Ser5 of the heptad repeat late in the process of transcription initiation (see Figure 6–16). This strategy ensures that the RNA molecule is efficiently capped as soon as its 5′ end emerges from the RNA polymerase. As the polymerase continues transcribing, its tail is extensively phosphorylated on the Ser2 positions by a kinase associated with the elongating polymerase and is eventually dephosphorylated at Ser5 positions. These further modifications attract splicing and 3′-end processing proteins to the moving polymerase, positioning them to act on the newly synthesized RNA as it emerges from the RNA polymerase. There are many RNA-processing enzymes, and not all travel with the polymerase. For RNA splicing, for example, the tail carries only a few critical components; once transferred to an RNA molecule, they serve as a nucleation site for the remaining components.

When RNA polymerase II finishes transcribing a gene, it is released from DNA, soluble phosphatases remove the phosphates on its tail, and it can reinitiate transcription. Only the dephosphorylated form of RNA polymerase II is competent to begin RNA synthesis at a promoter.

RNAs during transcription, in part because these polymerases lack a CTD. In the nucleus, the cap binds a protein complex called CBC (cap-binding complex), which, as we discuss in subsequent sections, helps the RNA to be properly processed and exported. The 5′-methyl cap also has an important role in the translation of mRNAs in the cytosol, as we discuss later in the chapter.

RNA Splicing Removes Intron Sequences from Newly Transcribed Pre-mRNAs

As discussed in Chapter 4, the protein coding sequences of eucaryotic genes are typically interrupted by noncoding intervening sequences (introns). Discovered in 1977, this feature of eucaryotic genes came as a surprise to scientists, who had been, until that time, familiar only with bacterial genes, which typically consist of a continuous stretch of coding DNA that is directly transcribed into mRNA. In marked contrast, eucaryotic genes were found to be broken up into small pieces of coding sequence (*expressed sequences* or **exons**) interspersed with much longer *intervening sequences* or **introns**; thus, the coding portion of a eucaryotic gene is often only a small fraction of the length of the gene (**Figure 6–25**).

Both intron and exon sequences are transcribed into RNA. The intron sequences are removed from the newly synthesized RNA through the process of **RNA splicing**. The vast majority of RNA splicing that takes place in cells functions in the production of mRNA, and our discussion of splicing focuses on this so-called precursor-mRNA (or pre-mRNA) splicing. Only after 5′ and 3′ end processing and splicing have taken place is such RNA termed mRNA.

Each splicing event removes one intron, proceeding through two sequential phosphoryl-transfer reactions known as transesterifications; these join two exons while removing the intron as a "lariat" (**Figure 6–26**). Since the number of high-energy phosphate bonds remains the same, these reactions could in

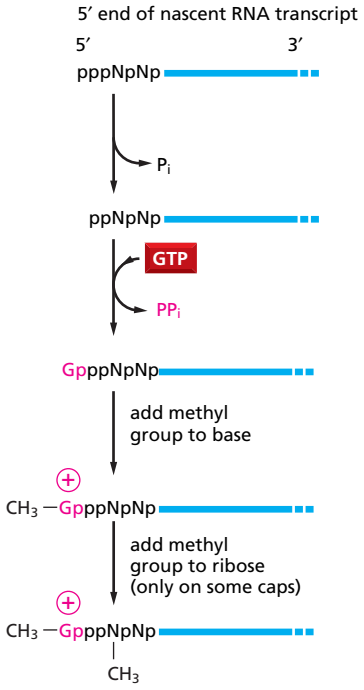


Figure 6–24 The reactions that cap the 5′ end of each RNA molecule synthesized by RNA polymerase II. The final cap contains a novel 5′-to-5′ linkage between the positively charged 7-methyl G residue and the 5′ end of the RNA transcript (see Figure 6–22B). The letter N represents any one of the four ribonucleotides, although the nucleotide that starts an RNA chain is usually a purine (an A or a G). (After A.J. Shatkin, *BioEssays* 7:275–277, 1987. With permission from ICSU Press.)

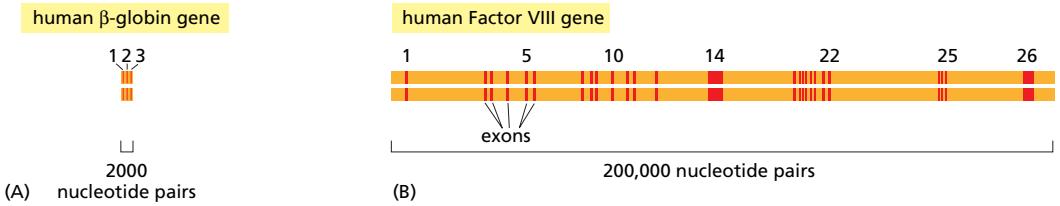


Figure 6–25 Structure of two human genes showing the arrangement of exons and introns. (A) The relatively small β-globin gene, which encodes one of the subunits of the oxygen-carrying protein hemoglobin, contains 3 exons (see also Figure 4–7). (B) The much larger Factor VIII gene contains 26 exons; it codes for a protein (Factor VIII) that functions in the blood-clotting pathway. The most prevalent form of hemophilia results from mutations in this gene.

principle take place without nucleoside triphosphate hydrolysis. However, the machinery that catalyzes pre-mRNA splicing is complex, consisting of 5 additional RNA molecules and as many as 200 proteins, and it hydrolyzes many ATP molecules per splicing event. This additional complexity ensures that splicing is accurate, while at the same time being flexible enough to deal with the enormous variety of introns found in a typical eucaryotic cell.

It may seem wasteful to remove large numbers of introns by RNA splicing. In attempting to explain why it occurs, scientists have pointed out that the exon–intron arrangement would seem to facilitate the emergence of new and useful proteins over evolutionary time scales. Thus, the presence of numerous introns in DNA allows genetic recombination to readily combine the exons of different genes (see p. 140), enabling genes for new proteins to evolve more easily by the combination of parts of preexisting genes. The observation, described in Chapter 3, that many proteins in present-day cells resemble patchworks composed from a common set of protein *domains*, supports this idea.

RNA splicing also has a present-day advantage. The transcripts of many eucaryotic genes (estimated at 75% of genes in humans) are spliced in more than one way, thereby allowing the same gene to produce a corresponding set of different proteins (**Figure 6–27**). Rather than being the wasteful process it may have seemed at first sight, RNA splicing enables eucaryotes to increase the already enormous coding potential of their genomes. We shall return to this idea again in this chapter and the next, but we first need to describe the cellular machinery that performs this remarkable task.

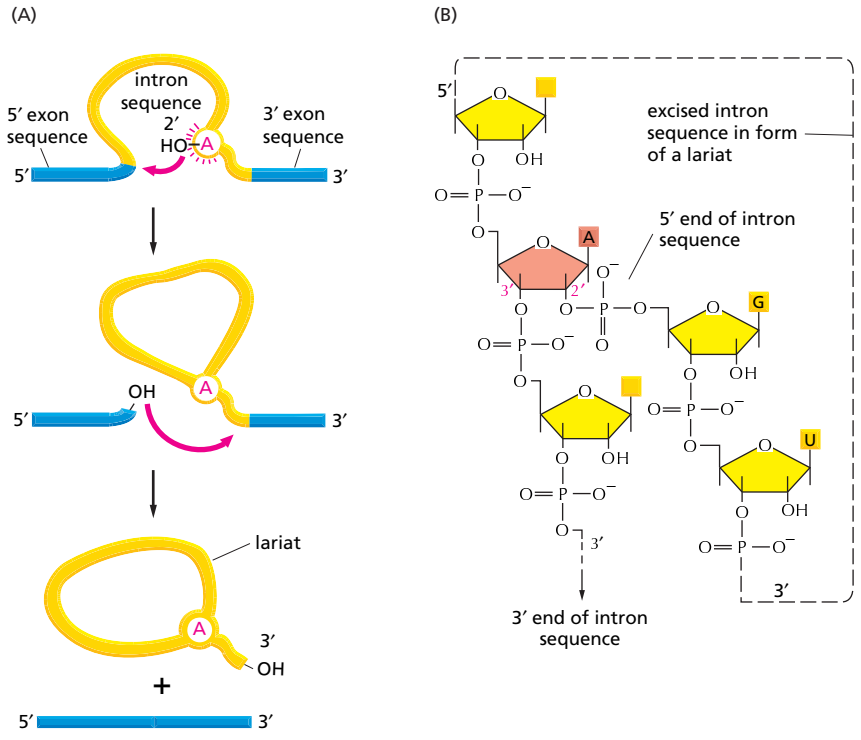


Figure 6–26 The pre-mRNA splicing reaction. (A) In the first step, a specific adenine nucleotide in the intron sequence (indicated in *red)* attacks the 5′ splice site and cuts the sugar-phosphate backbone of the RNA at this point. The cut 5′ end of the intron becomes covalently linked to the adenine nucleotide, as shown in detail in (B), thereby creating a loop in the RNA molecule. The released free 3′-OH end of the exon sequence then reacts with the start of the next exon sequence, joining the two exons together and releasing the intron sequence in the shape of a *lariat*. The two exon sequences thereby become joined into a continuous coding sequence; the released intron sequence is eventually degraded.

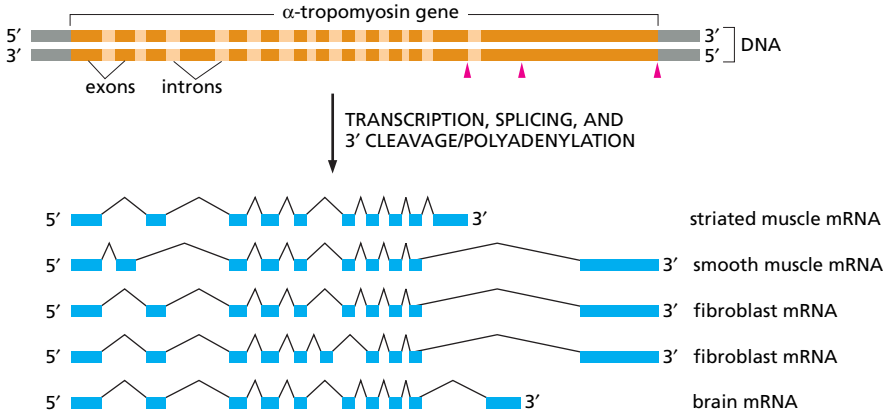


Figure 6–27 Alternative splicing of the α-tropomyosin gene from rat. α-Tropomyosin is a coiled-coil protein (see Figure 3–9) that regulates contraction in muscle cells. The primary transcript can be spliced in different ways, as indicated in the figure, to produce distinct mRNAs, which then give rise to variant proteins. Some of the splicing patterns are specific for certain types of cells. For example, the α-tropomyosin made in striated muscle is different from that made from the same gene in smooth muscle. The arrowheads in the top part of the figure mark the sites where cleavage and poly-A addition form the 3′ ends of the mature mRNAs.

Nucleotide Sequences Signal Where Splicing Occurs

The mechanism of pre-mRNA splicing shown in Figure 6–26 implies that the splicing machinery must recognize three portions of the precursor RNA molecule: the 5′ splice site, the 3′ splice site, and the branch point in the intron sequence that forms the base of the excised lariat. Not surprisingly, each site has a consensus nucleotide sequence that is similar from intron to intron and provides the cell with cues for where splicing is to take place (**Figure 6–28**). However, these consensus sequences are relatively short and can accommodate a high degree of sequence variability; as we shall see shortly, the cell incorporates additional types of information to ultimately choose exactly where, on each RNA molecule, splicing is to take place.

The high variability of the splicing consensus sequences presents a special challenge for scientists attempting to decipher genome sequences. Introns range in size from about 10 nucleotides to over 100,000 nucleotides, and choosing the precise borders of each intron is a difficult task even with the aid of powerful computers. The possibility of alternative splicing compounds the problem of predicting protein sequences solely from a genome sequence. This difficulty is one of the main barriers to identifying all of the genes in a complete genome sequence, and it is one of the primary reasons why we know only the approximate number of genes in the human genome.

RNA Splicing Is Performed by the Spliceosome

Unlike the other steps of mRNA production we have discussed, key steps in RNA splicing are performed by RNA molecules rather than proteins. Specialized RNA molecules recognize the nucleotide sequences that specify where splicing is to occur and also participate in the chemistry of splicing. These RNA molecules are relatively short (less than 200 nucleotides each), and there are five of them (U1, U2, U4, U5, and U6) involved in the major form of pre-mRNA splicing. Known as

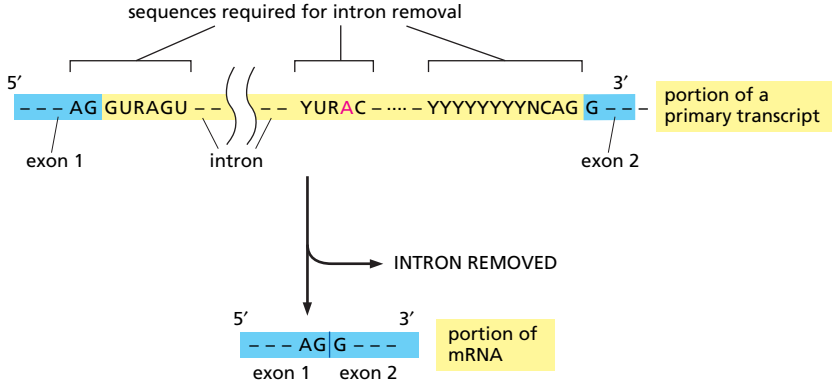
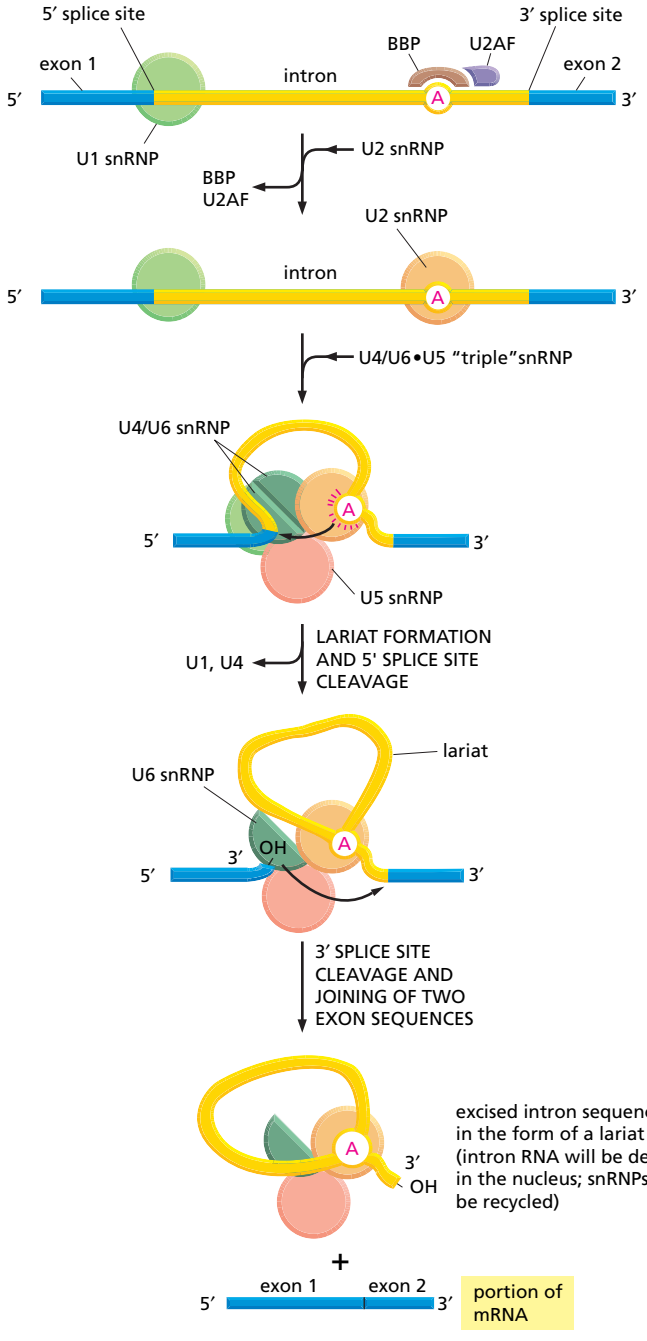


Figure 6–28 The consensus nucleotide sequences in an RNA molecule that signal the beginning and the end of most introns in humans. Only the three blocks of nucleotide sequences shown are required to remove an intron sequence; the rest of the intron can be occupied by any nucleotides. Here A, G, U, and C are the standard RNA nucleotides; R stands for purines (A or G); and Y stands for pyrimidines (C or U). The A highlighted in *red* forms the branch point of the lariat produced by splicing. Only the GU at the start of the intron and the AG at its end are invariant nucleotides in the splicing consensus sequences. Several different nucleotides can occupy the remaining positions (even the branch point A), although the indicated nucleotides are preferred. The distances along the RNA between the three splicing consensus sequences are highly variable; however, the distance between the branch point and 3′ splice junction is typically much shorter than that between the 5′ splice junction and the branch point.

snRNAs (**small nuclear RNAs**), each is complexed with at least seven protein subunits to form a snRNP (small nuclear ribonucleoprotein). These snRNPs form the core of the **spliceosome**, the large assembly of RNA and protein molecules that performs pre-mRNA splicing in the cell.

The spliceosome is a complex and dynamic machine. When studied *in vitro,* a few components of the spliceosome assemble on pre-mRNA and, as the splicing reaction proceeds, new components enter as those that have already performed their tasks are jettisoned (**Figure 6–29**). However, many scientists believe that, inside the cell, the spliceosome is a preexisting, loose assembly of all the components—capturing, splicing and releasing RNA as a coordinated unit, and undergoing extensive rearrangements each time a splice is made. During the splicing reaction, recognition of the 5′ splice junction, the branch-point site, and the 3′ splice junction is performed largely through base-pairing between the snRNAs and the consensus RNA sequences in the pre-mRNA substrate (**Figure**



The U1 snRNP forms base pairs with the 5′ splice junction (see Figure 6–30A) and the BBP (branch-point binding protein) and U2AF (U2 auxilliary factor) recognize the branch-point site.

The U2 snRNP displaces BBP and U2AF and forms base pairs with the branch-point site consensus sequence (see Figure 6–30B).

The U4/U6•U5 "triple" snRNP enters the reaction. In this triple snRNP, the U4 and U6 snRNAs are held firmly together by base-pair interactions. Subsequent rearrangements create the active site of the spliceosome and position the appropriate portions of the pre-mRNA substrate for the first phosphoryl-transferase reaction.

Several more RNA–RNA rearrangements occur that break apart the U4/U6 base pairs and allow the U6 snRNP to displace U1 at the 5′ splice junction (see Figure 6–30A) to form the active site for the second phosphoryl-transferase reaction, which completes the splice.

Figure 6–29 The pre-mRNA splicing mechanism. RNA splicing is catalyzed by an assembly of snRNPs (shown as *colored circles)* plus other proteins (most of which are not shown), which together constitute the spliceosome. The spliceosome recognizes the splicing signals on a pre-mRNA molecule, brings the two ends of the intron together, and provides the enzymatic activity for the two reaction steps (see Figure 6–26).

6–30). In the course of splicing, the spliceosome undergoes several shifts in which one set of base-pair interactions is broken and another is formed in its place. For example, U1 is replaced by U6 at the 5′ splice junction (see Figure 6–30A). This type of RNA–RNA rearrangement (in which the formation of one RNA–RNA interaction requires the disruption of another) occurs several times during the splicing reaction. It permits the checking and rechecking of RNA sequences before the chemical reaction is allowed to proceed, thereby increasing the accuracy of splicing.

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

Although ATP hydrolysis is not required for the chemistry of RNA splicing *per se*, it is required for the assembly and rearrangements of the spliceosome. Some of the additional proteins that make up the spliceosome use the energy of ATP hydrolysis to break existing RNA–RNA interactions to allow the formation of new ones. In fact, all the steps shown previously in Figure 6–29—except the association of BBP with the branch-point site and U1 snRNP with the 5′ splice site—require ATP hydrolysis and additional proteins. Each successful splice requires as many as 200 proteins, if we include those that form the snRNPs.

The ATP-requiring RNA–RNA rearrangements that take place in the spliceosome occur within the snRNPs themselves and between the snRNPs and the pre-mRNA substrate. One of the most important functions of these rearrangements is the creation of the active catalytic site of the spliceosome. The strategy of creating an active site only after the assembly and rearrangement of splicing components on a pre-mRNA substrate is a particularly effective way to prevent wayward splicing.

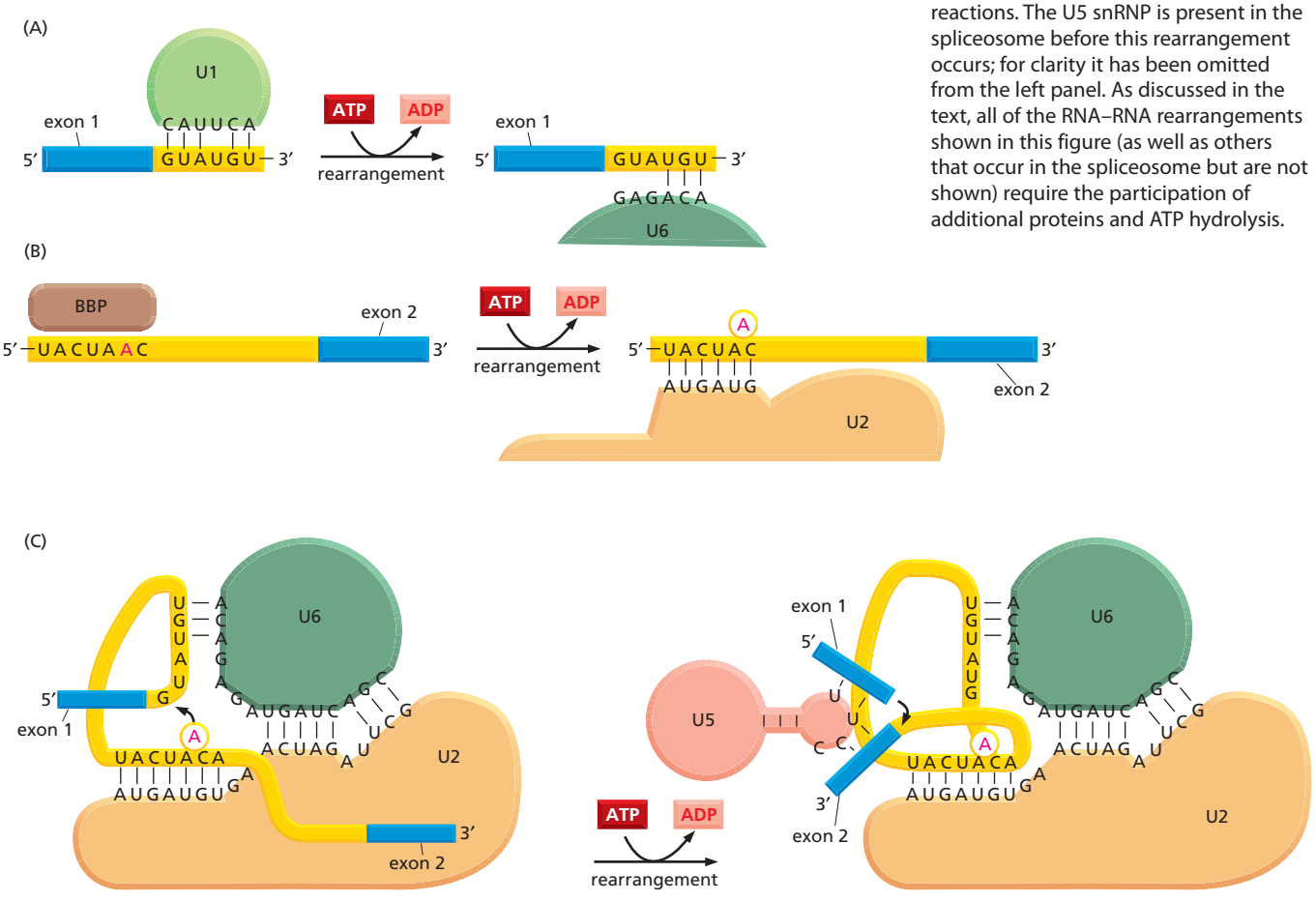


Figure 6–30 Several of the rearrangements that take place in the spliceosome during pre-mRNA splicing. Shown here are the details for the yeast *Saccharomyces cerevisiae*, in which the nucleotide sequences involved are slightly different from those in human cells. (A) The exchange of U1 snRNP for U6 snRNP occurs before the first phosphoryl-transfer reaction (see Figure 6–29). This exchange requires the 5′ splice site to be read by two different snRNPs, thereby increasing the accuracy of 5′ splice site selection by the spliceosome. (B) The branch-point site is first recognized by BBP and subsequently by U2 snRNP; as in (A), this "check and recheck" strategy provides increased accuracy of site selection. The binding of U2 to the branch point forces the appropriate adenine (in *red)* to be unpaired and thereby activates it for the attack on the 5′ splice site (see Figure 6–29). This, in combination with recognition by BBP, is the way in which the spliceosome accurately chooses the adenine that is ultimately to form the branch point. (C) After the first phosphoryl-transfer reaction *(left)* has occurred, a series of rearrangements brings the two exons into close proximity for the second phosphoryl-transfer reaction *(right)*. The snRNAs both position the reactants and provide (either all or in part) the catalytic sites for the two reactions. The U5 snRNP is present in the spliceosome before this rearrangement occurs; for clarity it has been omitted from the left panel. As discussed in the text, all of the RNA–RNA rearrangements shown in this figure (as well as others that occur in the spliceosome but are not shown) require the participation of additional proteins and ATP hydrolysis.

Perhaps the most surprising feature of the spliceosome is the nature of the catalytic site itself: it is largely (if not exclusively) formed by RNA molecules instead of proteins. In the last section of this chapter we discuss in general terms the structural and chemical properties of RNA that allow it to perform catalysis; here we need only consider that the U2 and U6 snRNAs in the spliceosome form a precise three-dimensional RNA structure that juxtaposes the 5′ splice site of the pre-mRNA with the branch-point site and probably performs the first transesterification reaction (see Figure 6–30C). In a similar way, the 5′ and 3′ splice junctions are brought together (an event requiring the U5 snRNA) to facilitate the second transesterification.

Once the splicing chemistry is completed, the snRNPs remain bound to the lariat. The disassembly of these snRNPs from the lariat (and from each other) requires another series of RNA–RNA rearrangements that require ATP hydrolysis, thereby returning the snRNAs to their original configuration so that they can be used again in a new reaction. At the completion of a splice, the spliceosome directs a set of proteins to bind to the mRNA near the position formerly occupied by the intron. Called the *exon junction complex (EJC),* these proteins mark the site of a successful splicing event and, as we shall see later in this chapter, influence the subsequent fate of the mRNA.

Other Properties of Pre-mRNA and Its Synthesis Help to Explain the Choice of Proper Splice Sites

As we have seen, intron sequences vary enormously in size, with some being in excess of 100,000 nucleotides. If splice-site selection were determined solely by the snRNPs acting on a preformed, protein-free RNA molecule, we would expect splicing mistakes—such as exon skipping and the use of "cryptic" splice sites—to be very common (**Figure 6–31**). The fidelity mechanisms built into the spliceosome, however, are supplemented by two additional strategies that increase the accuracy of splicing. The first is simply a consequence of the early stages of splicing occurring while the pre-mRNA molecules are being synthesized by RNA polymerase II. As transcription proceeds, the phosphorylated tail of RNA polymerase carries several components of the spliceosome (see Figure 6–23), and these components are transferred directly from the polymerase to the RNA as RNA is synthesized. This strategy helps the cell keep track of introns and exons: for example, the snRNPs that assemble at a 5′ splice site are initially presented with only a single 3′ splice site since the sites further downstream have not yet been synthesized. The coordination of transcription with splicing is especially important in preventing inappropriate exon skipping.

A strategy called "exon definition" is another way cells choose the appropriate splice sites. Exon size tends to be much more uniform than intron size, averaging about 150 nucleotide pairs across a wide variety of eucaryotic organisms (**Figure 6–32**). According to the exon definition idea, the splicing machinery initially seeks out the relatively homogenously sized exon sequences. As RNA synthesis proceeds, a group of additional components (most notably SR proteins, so-named because they contain a domain rich in serines and arginines) assemble on exon sequences and help to mark off each 3′ and 5′ splice site starting at the 5′ end of the RNA (**Figure 6–33**). These proteins, in turn, recruit U1 snRNA, which

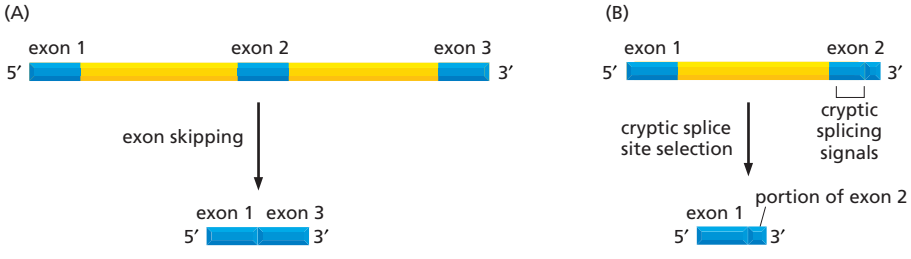


Figure 6–31 Two types of splicing errors. (A) Exon skipping. (B) Cryptic splice-site selection. Cryptic splicing signals are nucleotide sequences of RNA that closely resemble true splicing signals.

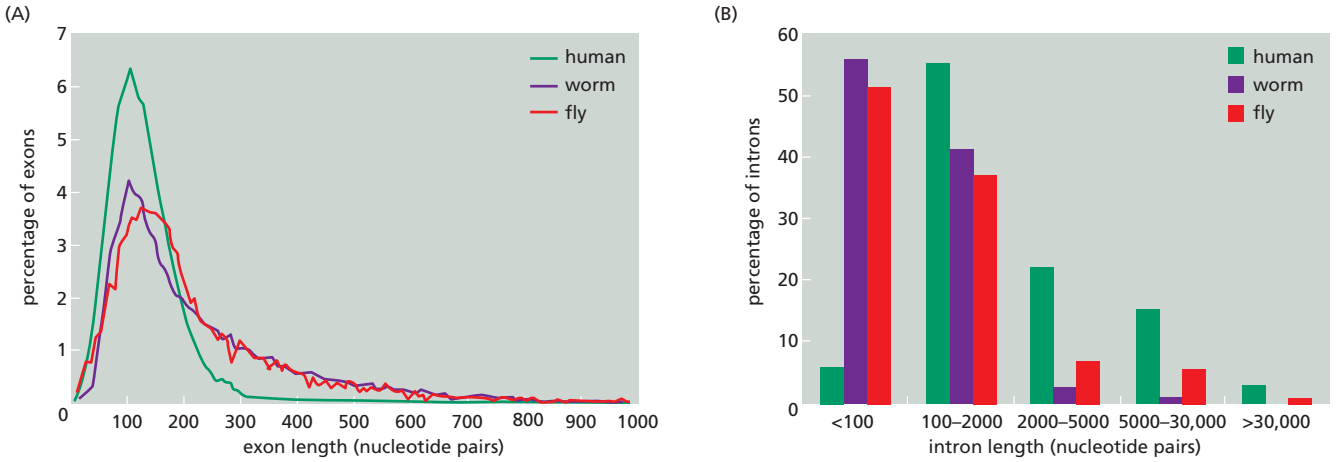


Figure 6–32 Variation in intron and exon lengths in the human, worm, and fly genomes. (A) Size distribution of exons. (B) Size distribution of introns. Note that exon length is much more uniform than intron length. (Adapted from International Human Genome Sequencing Consortium, *Nature* 409:860–921, 2001. With permission from Macmillan Publishers Ltd.)

marks the downstream exon boundary, and U2AF, which specifies the upstream one. By specifically marking the exons in this way and thereby taking advantage of the relatively uniform size of exons, the cell increases the accuracy with which it deposits the initial splicing components on the nascent RNA and thereby helps to avoid cryptic splice sites. How the SR proteins discriminate exon sequences from intron sequences is not understood in detail; however, it is known that some of the SR proteins bind preferentially to specific RNA sequences in exons, termed *splicing enhancers*. In principle, since any one of several different codons can be used to code for a given amino acid, there is freedom to adjust the exon nucleotide sequence so as to form a binding site for an SR protein, without necessarily affecting the amino acid sequence that the exon specifies.

Both the marking of exon and intron boundaries and the assembly of the spliceosome begin on an RNA molecule while it is still being elongated by RNA polymerase at its 3′ end. However, the actual chemistry of splicing can take place much later. This delay means that intron sequences are not necessarily removed from a pre-mRNA molecule in the order in which they occur along the RNA chain. It also means that, although spliceosome assembly is co-transcriptional, the splicing reactions sometimes occur posttranscriptionally—that is, after a complete pre-mRNA molecule has been made.

A Second Set of snRNPs Splice a Small Fraction of Intron Sequences in Animals and Plants

Simple eucaryotes such as yeasts have only one set of snRNPs that perform all pre-mRNA splicing. However, more complex eucaryotes such as flies, mammals, and plants have a second set of snRNPs that direct the splicing of a small fraction of their intron sequences. This minor form of spliceosome recognizes a different set of RNA sequences at the 5′ and 3′ splice junctions and at the branch point; it is called the *U12-type spliceosome* because of the involvement of the

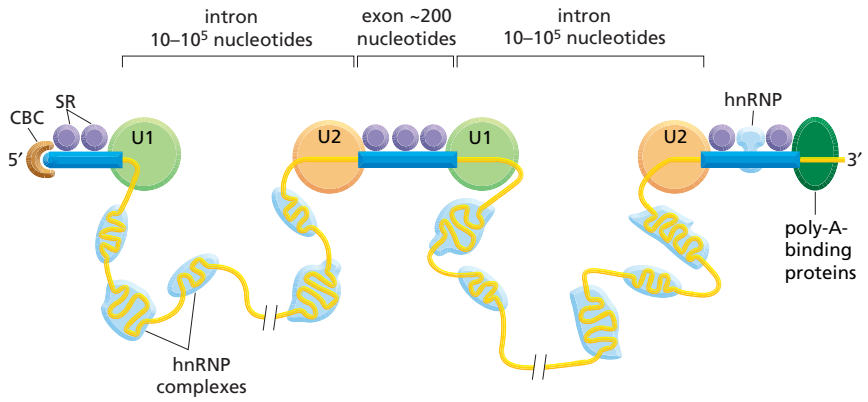
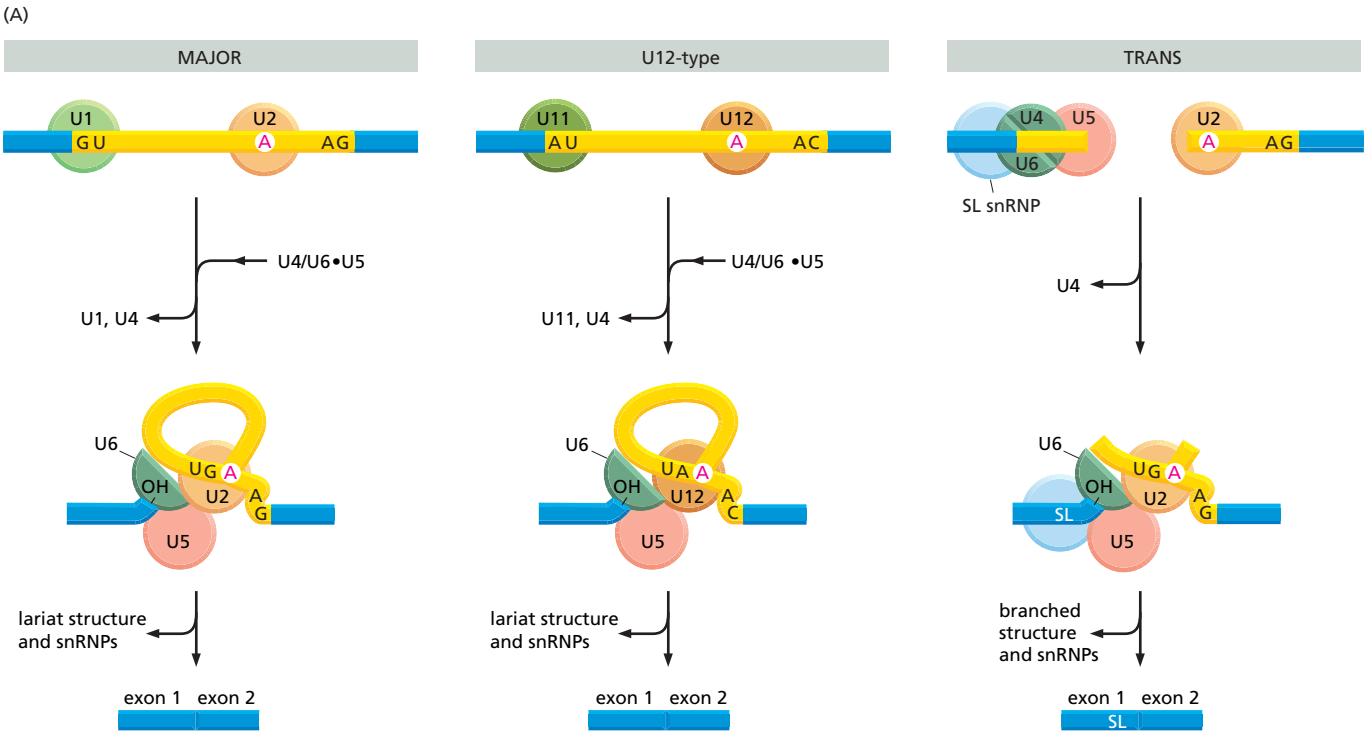


Figure 6–33 The exon definition idea. According to one proposal, SR proteins bind to each exon sequence in the pre-mRNA and thereby help to guide the snRNPs to the proper intron/exon boundaries. This demarcation of exons by the SR proteins occurs co-transcriptionally, beginning at the CBC (cap-binding complex) at the 5′ end. As indicated, the intron sequences in the pre-mRNA, which can be extremely long, are packaged into hnRNP (heterogeneous nuclear ribonucleoprotein) complexes that compact them into more manageable structures and perhaps mask cryptic splice sites. It has been proposed that hnRNP proteins may preferentially associate with intron sequences and that this preference may also help the spliceosome distinguish introns from exons. However, as shown, at least some hnRNP proteins also bind to exon sequences. (Adapted from R. Reed, *Curr. Opin. Cell Biol.* 12:340–345, 2000. With permission from Elsevier.)



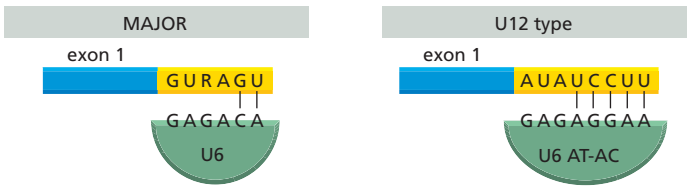


Figure 6–34 Outline of the mechanisms used for three types of RNA splicing. (A) Three types of spliceosomes. The major spliceosome *(left),* the U12-type spliceosome (*middle*), and the trans-spliceosome *(right)* are each shown at two stages of assembly. Introns removed by the U12-type spliceosome have a different set of consensus nucleotide sequences from those removed by the major spliceosome. In humans, it is estimated that 0.1% of introns are removed by the U12-type spliceosome. In trans-splicing, which does not occur in humans, the SL snRNP is consumed in the reaction because a portion of the SL snRNA becomes the first exon of the mature mRNA. (B) The major U6 snRNP and the U6 snRNP specific to the U12-type spliceosome both recognize the 5′ splice junction, but they do so through a different set of base-pair interactions. The sequences shown are from humans. (Adapted from Y.T. Yu et al., The RNA World, pp. 487–524. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press, 1999.)

U12 SnRNP (**Figure 6–34**A). Despite recognizing different nucleotide sequences, the snRNPs in this spliceosome make the same types of RNA–RNA interactions with the pre-mRNA and with each other as do the major snRNPs (Figure 6–34B). Although, as we have seen, components of the major spliceosomes travel with RNA polymerase II as it transcribes genes, this may not be the case for the U12 spliceosome. It is possible that U12-mediated splicing is thereby delayed, and this presents the cell with a way to co-regulate splicing of the several hundred genes whose expression requires this spliceosome. A number of mammalian mRNAs contain a mixture of introns, some removed by the major spliceosome and others by the minor spliceosome, and it has been proposed that this arrangement permits particularly complex patterns of alternative splicing to occur.

A few eucaryotic organisms exhibit a particular variation on splicing, called **trans-splicing**. These organisms include the single-celled trypanosomes—protozoans that cause African sleeping sickness in humans—and the model multicellular organism, the nematode worm. In trans-splicing, exons from two separate RNA transcripts are spliced together to form a mature mRNA molecule (see Figure 6–34A). Trypanosomes produce all of their mRNAs in this way, whereas trans-splicing accounts for only about 1% of nematode mRNAs. In both cases, a single exon is spliced onto the 5′ end of many different RNA transcripts produced by the cell; in this way, all of the products of trans-splicing have the same 5′ exon and different 3′ exons. Many of the same snRNPs that function in conventional splicing are used in this reaction, although trans-splicing uses a unique snRNP (called the SL RNP) that brings in the common exon (see Figure 6–34).

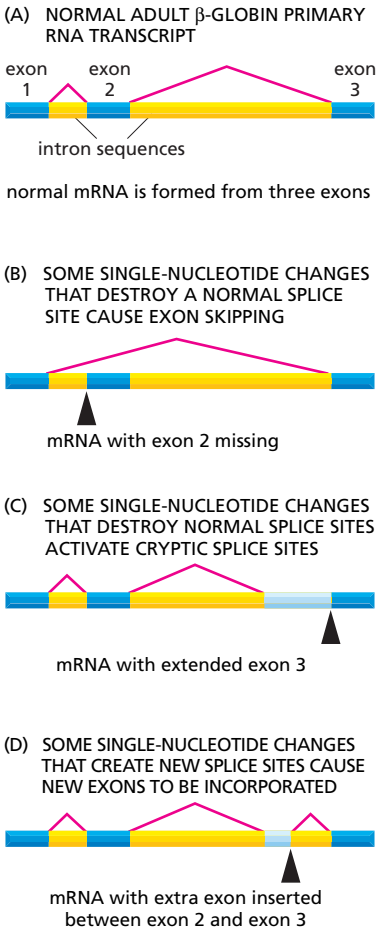


Figure 6–35 Abnormal processing of the β-globin primary RNA transcript in humans with the disease β thalassemia. In the examples shown, the disease is caused by splice-site mutations (*black arrowheads*) found in the genomes of affected patients. The *dark blue boxes* represent the three normal exon sequences; the *red lines* indicate the 5′ and 3′ splice sites. The *light blue boxes* depict new nucleotide sequences included in the final mRNA molecule as a result of the mutation. Note that when a mutation leaves a normal splice site without a partner, an exon is skipped or one or more abnormal cryptic splice sites nearby is used as the partner site, as in (C) and (D). (Adapted in part from S.H. Orkin, in The Molecular Basis of Blood Diseases [G. Stamatoyannopoulos et al., eds.], pp. 106–126. Philadelphia: Saunders, 1987.)

We do not know why even a few organisms use trans-splicing; however, it is thought that the common 5′ exon may aid in the translation of the mRNA. Thus, the mRNAs produced by trans-splicing in nematodes seem to be translated with especially high efficiency.

RNA Splicing Shows Remarkable Plasticity

We have seen that the choice of splice sites depends on such features of the pre-mRNA transcript as the affinity of the three signals on the RNA (the 5′ and 3′ splice junctions and the branch point) for the splicing machinery, the co-transcriptional assembly of the spliceosome, and the "bookkeeping" that underlies exon definition. We do not know how accurate splicing normally is because, as we see later, there are several quality control systems that rapidly destroy mRNAs whose splicing goes awry. However, we do know that, compared with other steps in gene expression, splicing is unusually flexible. For example, a mutation in a nucleotide sequence critical for splicing of a particular intron does not necessarily prevent splicing of that intron altogether. Instead, the mutation typically creates a new pattern of splicing (**Figure 6–35**). Most commonly, an exon is simply skipped (Figure 6–35B). In other cases, the mutation causes a cryptic splice junction to be efficiently used (Figure 6–35C). Apparently, the splicing machinery has evolved to pick out the best possible pattern of splice junctions, and if the optimal one is damaged by mutation, it will seek out the next best pattern, and so on. This flexibility in the process of RNA splicing suggests that changes in splicing patterns caused by random mutations have been an important pathway in the evolution of genes and organisms.

The plasticity of RNA splicing also means that the cell can regulate the pattern of RNA splicing. Earlier in this section we saw that alternative splicing can give rise to different proteins from the same gene. Some examples of alternative splicing are constitutive; that is, the alternatively spliced mRNAs are produced continuously by cells of an organism. However, in many cases, the cell regulates the splicing patterns so that different forms of the protein are produced at different times and in different tissues (see Figure 6–27). In Chapter 7 we return to this issue to discuss some specific examples of regulated RNA splicing.

Spliceosome-Catalyzed RNA Splicing Probably Evolved from Self-splicing Mechanisms

When the spliceosome was first discovered, it puzzled molecular biologists. Why do RNA molecules instead of proteins perform important roles in splice site recognition and in the chemistry of splicing? Why is a lariat intermediate used rather than the apparently simpler alternative of bringing the 5′ and 3′ splice sites together in a single step, followed by their direct cleavage and rejoining? The answers to these questions reflect the way in which the spliceosome is believed to have evolved.

As discussed briefly in Chapter 1 (and in more detail in the final section of this chapter), it is likely that early cells used RNA molecules rather than proteins as their major catalysts and that they stored their genetic information in RNA rather than in DNA sequences. RNA-catalyzed splicing reactions presumably had important roles in these early cells. As evidence, some *self-splicing RNA* introns (that is, intron sequences in RNA whose splicing out can occur in the absence of proteins or any other RNA molecules) remain today—for example, in the nuclear rRNA genes of the ciliate *Tetrahymena*, in a few bacteriophage T4 genes, and in some mitochondrial and chloroplast genes.

A self-splicing intron sequence can be identified in a test tube by incubating a pure RNA molecule that contains the intron sequence and observing the splicing reaction. Two major classes of self-splicing intron sequences can be distinguished in this way. *Group I intron sequences* begin the splicing reaction by binding a G nucleotide to the intron sequence; this G is thereby activated to form the attacking group that will break the first of the phosphodiester bonds cleaved during splicing (the bond at the 5′ splice site). In *group II intron sequences,* an especially reactive A residue in the intron sequence is the attacking group, and a lariat intermediate is generated. Otherwise the reaction pathways for the two types of self-splicing intron sequences are the same. Both are presumed to represent vestiges of very ancient mechanisms (**Figure 6–36**).

For both types of self-splicing reactions, the nucleotide sequence of the intron is critical; the intron RNA folds into a specific three-dimensional structure, which brings the 5′ and 3′ splice junctions together and provides precisely positioned reactive groups to perform the chemistry (see Figure 6–6C). Because the chemistries of their splicing reactions are so similar, it has been proposed that the pre-mRNA splicing mechanism of the spliceosome evolved from group

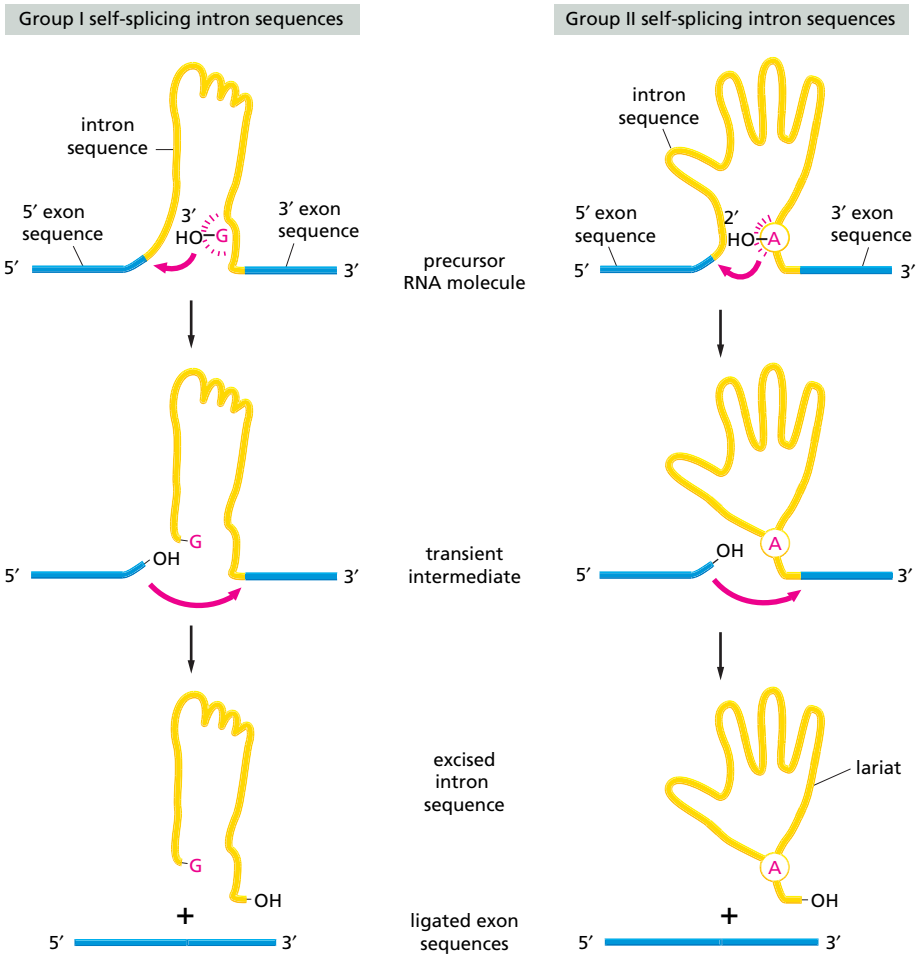


Figure 6–36 The two known classes of self-splicing intron sequences. The figure emphasizes the similarities between the two mechanisms. Both are normally aided by proteins in the cell that speed up the reaction, but the catalysis is nevertheless mediated by the RNA in the intron sequence. The group I intron sequences bind a free G nucleotide to a specific site on the RNA to initiate splicing, while the group II intron sequences use an especially reactive A nucleotide in the intron sequence itself for the same purpose. Both types of self-splicing reactions require the intron to be folded into a highly specific three-dimensional structure that provides the catalytic activity for the reaction (see Figure 6–6). The mechanism used by group II intron sequences releases the intron as a lariat structure and closely resembles the pathway of pre-mRNA splicing catalyzed by the spliceosome (compare with Figure 6–29). The spliceosome performs most RNA splicing in eucaryotic cells, and self-splicing RNAs represent unusual cases. (Adapted from T.R. Cech, *Cell* 44:207–210, 1986. With permission from Elsevier.)

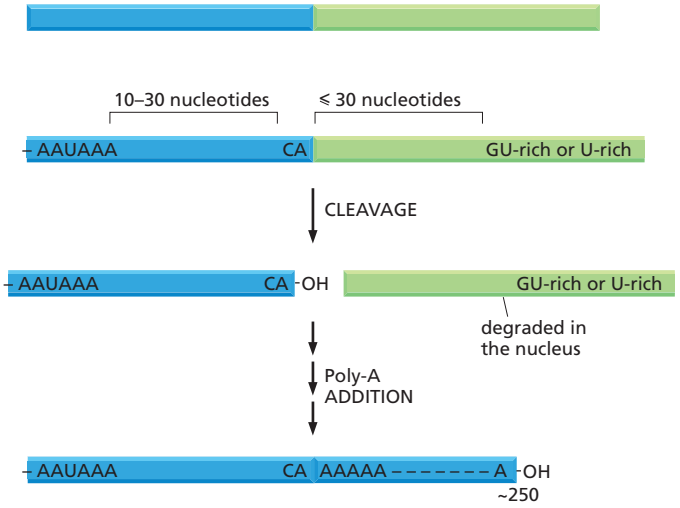


Figure 6–37 Consensus nucleotide sequences that direct cleavage and polyadenylation to form the 3′ end of a eucaryotic mRNA. These sequences are encoded in the genome; specific proteins recognize them after they are transcribed into RNA. The hexamer AAUAAA is bound by CPSF, the GU-rich element beyond the cleavage site is bound by CstF (see Figure 6–38), and the CA sequence is bound by a third factor required for the cleavage step. Like other consensus nucleotide sequences discussed in this chapter (see Figure 6–12), the sequences shown in the figure represent a variety of individual cleavage and polyadenylation signals.

II self-splicing. According to this idea, when the spliceosomal snRNPs took over the structural and chemical roles of the group II introns, the strict sequence constraints on intron sequences would have disappeared, thereby permitting a vast expansion in the number of different RNAs that could be spliced.

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

As previously explained, the 5′ end of the pre-mRNA produced by RNA polymerase II is capped almost as soon as it emerges from the RNA polymerase. Then, as the polymerase continues its movement along a gene, the spliceosome assembles on the RNA and delineates the intron and exon boundaries. The long C-terminal tail of the RNA polymerase coordinates these processes by transferring capping and splicing components directly to the RNA as it emerges from the enzyme. We see in this section that, as RNA polymerase II reaches the end of a gene, a similar mechanism ensures that the 3′ end of the pre-mRNA is appropriately processed.

As might be expected, the position of the 3′ end of each mRNA molecule is ultimately specified by a signal encoded in the genome (**Figure 6–37**). These signals are transcribed into RNA as the RNA polymerase II moves through them, and they are then recognized (as RNA) by a series of RNA-binding proteins and RNA-processing enzymes (**Figure 6–38**). Two multisubunit proteins, called CstF (cleavage stimulation factor) and CPSF (cleavage and polyadenylation specificity factor), are of special importance. Both of these proteins travel with the RNA polymerase tail and are transferred to the 3′-end processing sequence on an RNA molecule as it emerges from the RNA polymerase.

Once CstF and CPSF bind to specific nucleotide sequences on the emerging RNA molecule, additional proteins assemble with them to create the 3′ end of the mRNA. First, the RNA is cleaved (see Figure 6–38). Next an enzyme called poly-A polymerase (PAP) adds, one at a time, approximately 200 A nucleotides to the 3′ end produced by the cleavage. The nucleotide precursor for these additions is ATP, and the same type of 5′-to-3′ bonds are formed as in conventional RNA synthesis (see Figure 6–4). Unlike the usual RNA polymerases, poly-A polymerase does not require a template; hence the poly-A tail of eucaryotic mRNAs

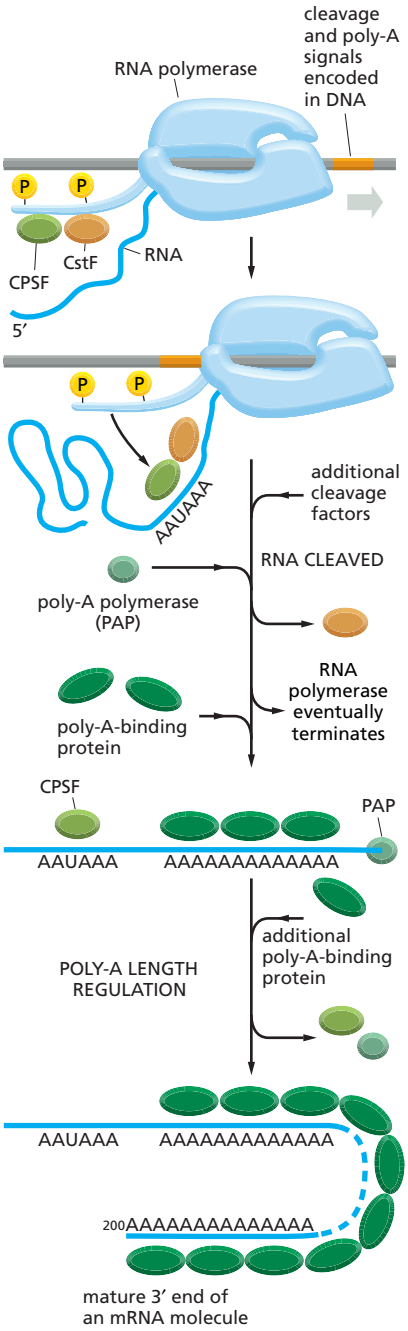


Figure 6–38 Some of the major steps in generating the 3′ end of a eucaryotic mRNA. This process is much more complicated than the analogous process in bacteria, where the RNA polymerase simply stops at a termination signal and releases both the 3′ end of its transcript and the DNA template (see Figure 6–11).

is not directly encoded in the genome. As the poly-A tail is synthesized, proteins called poly-A-binding proteins assemble onto it and, by a poorly understood mechanism, determine the final length of the tail. Some poly-A-binding proteins remain bound to the poly-A tail as the mRNA travels from the nucleus to the cytosol and they help to direct the synthesis of a protein on the ribosome, as we see later in this chapter.

After the 3′ end of a eucaryotic pre-mRNA molecule has been cleaved, the RNA polymerase II continues to transcribe, in some cases for hundreds of nucleotides. But the polymerase soon releases its grip on the template and transcription terminates. After 3′-end cleavage has occurred, the newly synthesized RNA that emerges from the polymerases lacks a 5′ cap; this unprotected RNA is rapidly degraded by a 5′ → 3′ exonuclease, which is carried along on the polymerase tail. Apparently, it is this RNA degradation that eventually causes the RNA polymerase to dissociate from the DNA.

Mature Eucaryotic mRNAs Are Selectively Exported from the Nucleus

We have seen how eucaryotic pre-mRNA synthesis and processing take place in an orderly fashion within the cell nucleus. However, these events create a special problem for eucaryotic cells, especially those of complex organisms where the introns are vastly longer than the exons. Of the pre-mRNA that is synthesized, only a small fraction—the mature mRNA—is of further use to the cell. The rest—excised introns, broken RNAs, and aberrantly processed pre-mRNAs—is not only useless but potentially dangerous. How, then, does the cell distinguish between the relatively rare mature mRNA molecules it wishes to keep and the overwhelming amount of debris from RNA processing?

The answer is that, as an RNA molecule is processed, it loses certain proteins and acquires others, thereby signifying the successful completion of each of the different steps. For example, we have seen that acquisition of the cap-binding complexes, the exon junction complexes, and the poly-A-binding proteins mark the completion of capping, splicing, and poly-A addition, respectively. A properly completed mRNA molecule is also distinguished by the proteins it lacks. For example, the presence of a snRNP would signify incomplete or aberrant splicing. Only when the proteins present on an mRNA molecule collectively signify that processing was successfully completed is the mRNA exported from the nucleus into the cytosol, where it can be translated into protein. Improperly processed mRNAs, and other RNA debris are retained in the nucleus, where they are eventually degraded by the nuclear **exosome**, a large protein complex whose interior is rich in 3′-to-5′ RNA exonucleases. Eucaryotic cells thus export only useful RNA molecules to the cytoplasm, while debris is disposed of in the nucleus.

Of all the proteins that assemble on pre-mRNA molecules as they emerge from transcribing RNA polymerases, the most abundant are the hnRNPs (heterogeneous nuclear ribonuclear proteins) (see Figure 6–33). Some of these proteins (there are approximately 30 of them in humans) unwind the hairpin helices from the RNA so that splicing and other signals on the RNA can be read more easily. Others preferentially package the RNA contained in the very long intron sequences typically found in genes of complex organisms. They may therefore play an important role in distinguishing mature mRNA from the debris left over from RNA processing.

Successfully processed mRNAs are guided through the **nuclear pore complexes** (NPCs)—aqueous channels in the nuclear membrane that directly connect the nucleoplasm and cytosol (**Figure 6–39**). Small molecules (less than 50,000 daltons) can diffuse freely through these channels. However, most of the macromolecules in cells, including mRNAs complexed with proteins, are far too large to pass through the channels without a special process. The cell uses energy to actively transport such macromolecules in both directions through the nuclear pore complexes.

As explained in detail in Chapter 12, macromolecules are moved through nuclear pore complexes by *nuclear transport receptors,* which, depending on the

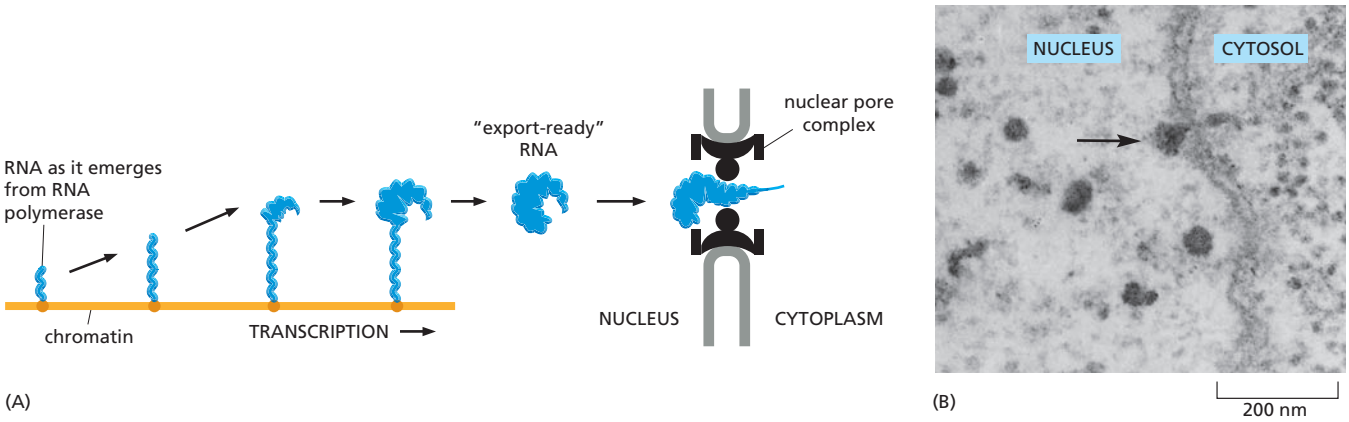


Figure 6–39 Transport of a large mRNA molecule through the nuclear pore complex. (A) The maturation of an mRNA molecule as it is synthesized by RNA polymerase and packaged by a variety of nuclear proteins. This drawing of an unusually abundant RNA, called the Balbiani Ring mRNA, is based on EM micrographs such as that shown in (B). Balbiani Rings are found in the cells of certain insects. (A, adapted from B. Daneholt, *Cell* 88:585–588, 1997. With permission from Elsevier; B, from B.J. Stevens and H. Swift, *J. Cell Biol*. 31:55–77, 1966. With permission from The Rockefeller University Press.)

identity of the macromolecule, escort it from the nucleus to the cytoplasm or vice versa. For mRNA export to occur, a specific nuclear transport receptor must be loaded onto the mRNA, a step that, at least in some organisms, takes place in concert with 3′ cleavage and polyadenylation. Once it helps to move an RNA molecule through the nuclear pore complex, the transport receptor dissociates from the mRNA, re-enters the nucleus, and exports a new mRNA molecule (**Figure 6–40**).

The export of mRNA–protein complexes from the nucleus can be observed with the electron microscope for the unusually abundant mRNA of the insect Balbiani Ring genes. As these genes are transcribed, the newly formed RNA is seen to be packaged by proteins, including hnRNPs, SR proteins, and components of the spliceosome. This protein–RNA complex undergoes a series of structural transitions, probably reflecting RNA processing events, culminating in a curved fiber (see Figure 6–39). This curved fiber moves through the nucleoplasm and enters the nuclear pore complex (with its 5′ cap proceeding first), and it then undergoes another series of structural transitions as it moves through the

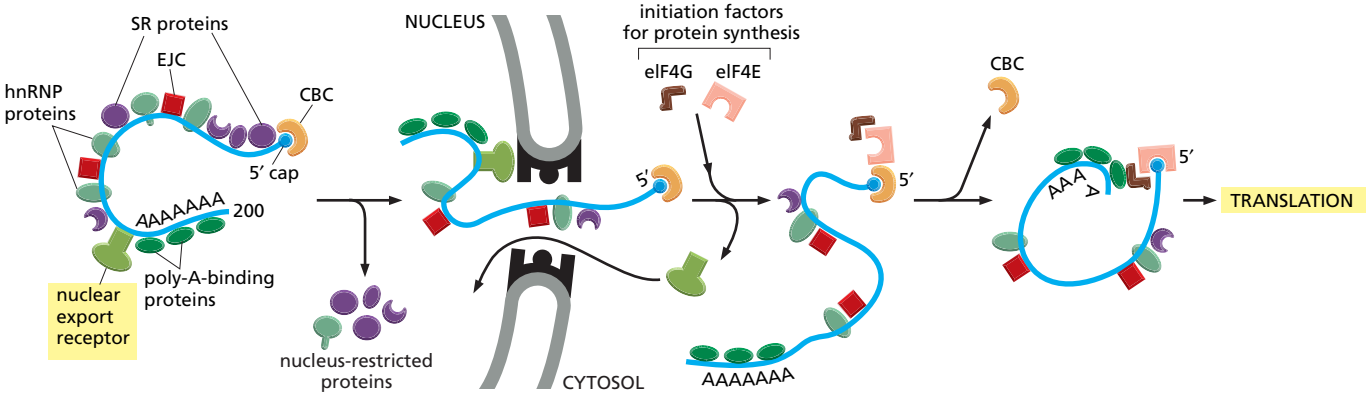


Figure 6–40 Schematic illustration of an "export-ready" mRNA molecule and its transport through the nuclear pore. As indicated, some proteins travel with the mRNA as it moves through the pore, whereas others remain in the nucleus. The nuclear export receptor for mRNAs is a complex of proteins that is deposited when the mRNA has been correctly spliced and polyadenylated. When the mRNA is exported to the cytosol, the export receptor dissociates from the mRNA and is re-imported into the nucleus, where it can be used again. Just after it leaves the nucleus, and before it loses the cap-binding complex (CBC) the mRNA is subjected to a final check, called *nonsense-mediated decay*, which is described later in the chapter. Once it passes this test the mRNA continues to shed previously bound proteins and acquire new ones before it is efficiently translated into protein. EJC, exon junction complex.

pore. These and other observations reveal that the pre-mRNA–protein and mRNA–protein complexes are dynamic structures that gain and lose numerous specific proteins during RNA synthesis, processing, and export (see Figure 6–40).

As we have seen, some of these proteins mark the different stages of mRNA maturation; other proteins deposited on the mRNA while it is still in the nucleus can affect the fate of the RNA after it is transported to the cytosol. Thus, the stability of an mRNA in the cytosol, the efficiency with which it is translated into protein, and its ultimate destination in the cytosol can all be determined by proteins acquired in the nucleus that remain bound to the RNA after it leaves the nucleus. We will discuss these issues in Chapter 7 when we turn to the post-transcriptional control of gene expression.

We have seen that RNA synthesis and processing are closely coupled in the cell, and it might be expected that export from the nucleus is somehow integrated with these two processes. Although the Balbiani Ring RNAs can be seen to move through the nucleoplasm and out through the nuclear pores, other mRNAs appear to be synthesized and processed in close proximity to nuclear pore complexes. In these cases, which may represent the majority of eucaryotic genes, mRNA synthesis, processing, and transport all appear to be tightly coupled; the mRNA can thus be viewed as emerging from the nuclear pore as a newly manufactured car might emerge from an assembly line. Later in this chapter, we will see that the cell performs an additional quality-control check on each mRNA before it is allowed to be efficiently translated into protein.

Before discussing what happens to mRNAs after they leave the nucleus, we briefly consider how the synthesis and processing of noncoding RNA molecules occurs. Although there are many other examples, our discussion focuses on the rRNAs that are critically important for the translation of mRNAs into protein.

Many Noncoding RNAs Are Also Synthesized and Processed in the Nucleus

A few percent of the dry weight of a mammalian cell is RNA; of that, only about 3–5% is mRNA. A fraction of the remainder represents intron sequences before they have been degraded, but the bulk of the RNA in cells performs structural and catalytic functions (see Table 6–1, p. 336). The most abundant RNAs in cells are the ribosomal RNAs (rRNAs), constituting approximately 80% of the RNA in rapidly dividing cells. As discussed later in this chapter, these RNAs form the core of the ribosome. Unlike bacteria—in which a single RNA polymerase synthesizes all RNAs in the cell—eucaryotes have a separate, specialized polymerase, RNA polymerase I, that is dedicated to producing rRNAs. RNA polymerase I is similar structurally to the RNA polymerase II discussed previously; however, the absence of a C-terminal tail in polymerase I helps to explain why its transcripts are neither capped nor polyadenylated. As mentioned earlier, this difference helps the cell distinguish between noncoding RNAs and mRNAs.

Because multiple rounds of translation of each mRNA molecule can provide an enormous amplification in the production of protein molecules, many of the proteins that are very abundant in a cell can be synthesized from genes that are present in a single copy per haploid genome. In contrast, the RNA components of the ribosome are final gene products, and a growing mammalian cell must synthesize approximately 10 million copies of each type of ribosomal RNA in each cell generation to construct its 10 million ribosomes. The cell can produce adequate quantities of ribosomal RNAs only because it contains multiple copies of the **rRNA genes** that code for ribosomal RNAs (**rRNAs**). Even *E. coli* needs seven copies of its rRNA genes to meet the cell's need for ribosomes. Human cells contain about 200 rRNA gene copies per haploid genome, spread out in small clusters on five different chromosomes (see Figure 4–11), while cells of the frog *Xenopus* contain about 600 rRNA gene copies per haploid genome in a single cluster on one chromosome (**Figure 6–41**).

There are four types of eucaryotic rRNAs, each present in one copy per ribosome. Three of the four rRNAs (18S, 5.8S, and 28S) are made by chemically modifying and cleaving a single large precursor rRNA (**Figure 6–42**); the fourth (5S

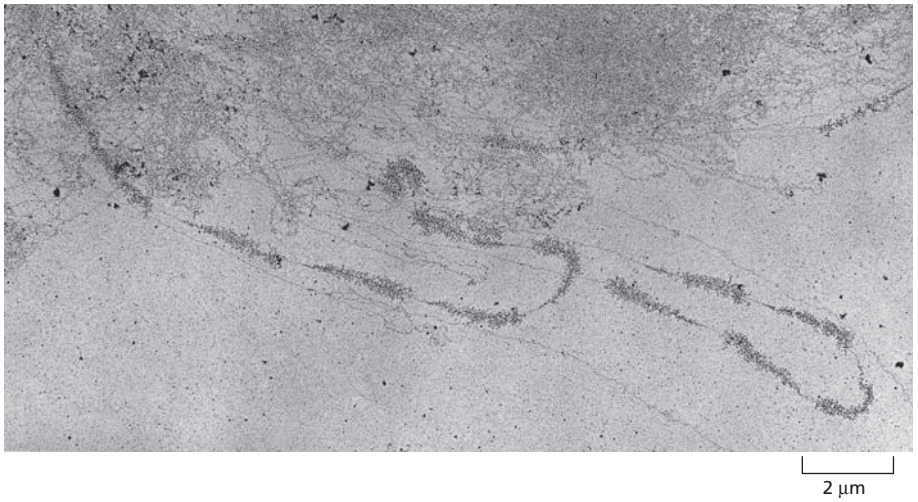


Figure 6–41 Transcription from tandemly arranged rRNA genes, as seen in the electron microscope. The pattern of alternating transcribed gene and nontranscribed spacer is readily seen. A higher-magnification view of rRNA genes is shown in Figure 6–9. (From V.E. Foe, *Cold Spring Harbor Symp. Quant. Biol*. 42:723–740, 1978. With permission from Cold Spring Harbor Laboratory Press.)

RNA) is synthesized from a separate cluster of genes by a different polymerase, RNA polymerase III, and does not require chemical modification.

Extensive chemical modifications occur in the 13,000-nucleotide-long precursor rRNA before the rRNAs are cleaved out of it and assembled into ribosomes. These include about 100 methylations of the 2′-OH positions on nucleotide sugars and 100 isomerizations of uridine nucleotides to pseudouridine (**Figure 6–43**A). The functions of these modifications are not understood in detail, but many probably aid in the folding and assembly of the final rRNAs and some may subtly alter the function of ribosomes. Each modification is made at a specific position in the precursor rRNA. These positions are specified by about 150 "guide RNAs," which position themselves through base-pairing to the precursor rRNA and thereby bring an RNA-modifying enzyme to the appropriate position (Figure 6–43B). Other guide RNAs promote cleavage of the precursor rRNAs into the mature rRNAs, probably by causing conformational changes in the precursor rRNA that expose these sites to nucleases. All of these guide RNAs are members of a large class of RNAs called **small nucleolar RNAs** (or **snoRNAs**), so named because these RNAs perform their functions in a subcompartment of the nucleus called the nucleolus. Many snoRNAs are encoded in the introns of other genes, especially those encoding ribosomal proteins. They are therefore synthesized by RNA polymerase II and processed from excised intron sequences.

Recently several snoRNA-like RNAs have been identified that are synthesized only in cells of the brain. These are believed to direct the modification of mRNAs, instead of rRNAs, and are likely to represent a new, but poorly understood, type of gene regulatory mechanism.

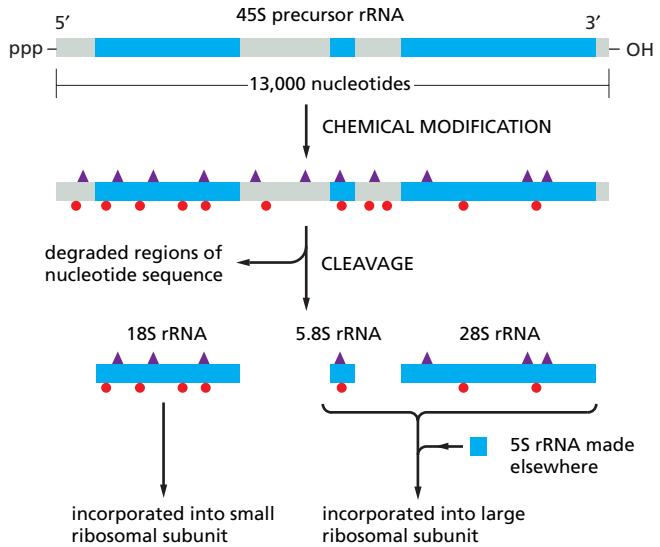


Figure 6–42 The chemical modification and nucleolytic processing of a eucaryotic 45S precursor rRNA molecule into three separate ribosomal RNAs. Two types of chemical modifications (color-coded as indicated in Figure 6–43) are made to the precursor rRNA before it is cleaved. Nearly half of the nucleotide sequences in this precursor rRNA are discarded and degraded in the nucleus. The rRNAs are named according to their "S" values, which refer to their rate of sedimentation in an ultracentrifuge. The larger the S value, the larger the rRNA.

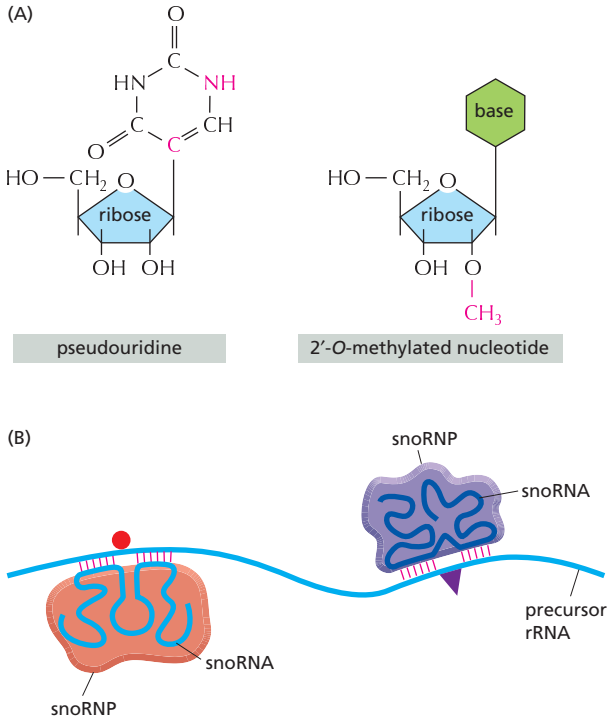
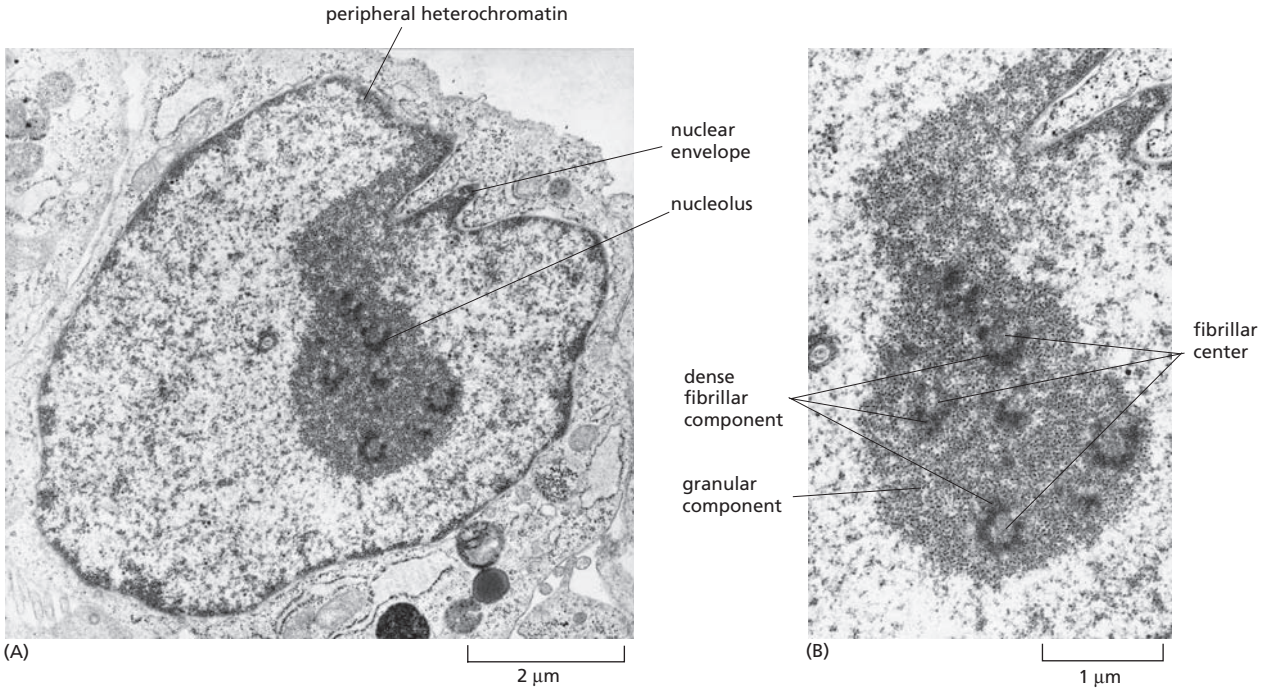


Figure 6–43 Modifications of the precursor rRNA by guide RNAs. (A) Two prominent covalent modifications occur after rRNA synthesis; the differences from the initially incorporated nucleotide are indicated by *red* atoms. Pseudouridine is an isomer of uridine; the base has been "rotated" relative to the sugar. (B) As indicated, snoRNAs determine the sites of modification by base-pairing to complementary sequences on the precursor rRNA. The snoRNAs are bound to proteins, and the complexes are called snoRNPs. snoRNPs contain both the guide sequences and the enzymes that modify the rRNA.

The Nucleolus Is a Ribosome-Producing Factory

The nucleolus is the most obvious structure seen in the nucleus of a eucaryotic cell when viewed in the light microscope. Consequently, it was so closely scrutinized by early cytologists that an 1898 review could list some 700 references. We now know that the nucleolus is the site for the processing of rRNAs and their assembly into ribosome subunits. Unlike many of the major organelles in the cell, the nucleolus is not bound by a membrane (**Figure 6–44**); instead, it is a large aggregate of macromolecules, including the rRNA genes themselves, precursor rRNAs, mature rRNAs, rRNA-processing enzymes, snoRNPs, ribosomal proteins and partly assembled ribosomes. The close association of all these components presumably allows the assembly of ribosomes to occur rapidly and smoothly.

Figure 6–44 Electron micrograph of a thin section of a nucleolus in a human fibroblast, showing its three distinct zones. (A) View of entire nucleus. (B) High-power view of the nucleolus. It is believed that transcription of the rRNA genes takes place between the fibrillar center and the dense fibrillar component and that processing of the rRNAs and their assembly into the two subunits of the ribosome proceeds outward from the dense fibrillar component to the surrounding granular components. (Courtesy of E.G. Jordan and J. McGovern.)



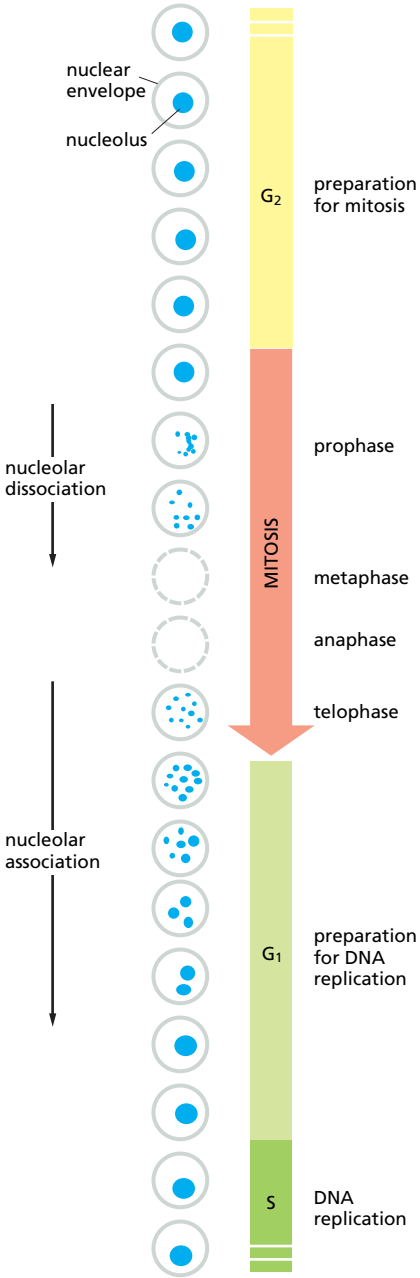


Figure 6–45 Changes in the appearance of the nucleolus in a human cell during the cell cycle. Only the cell nucleus is represented in this diagram. In most eucaryotic cells the nuclear envelope breaks down during mitosis, as indicated by the dashed circles.

Various types of RNA molecules play a central part in the chemistry and structure of the nucleolus, suggesting that it may have evolved from an ancient structure present in cells dominated by RNA catalysis. In present-day cells, the rRNA genes also have an important role in forming the nucleolus. In a diploid human cell, the rRNA genes are distributed into 10 clusters, located near the tips of five different chromosome pairs (see Figure 4–11). During interphase these 10 chromosomes contribute DNA loops (containing the rRNA genes) to the nucleolus; in M-phase, when the chromosomes condense, the nucleolus disappears. Finally, in the telophase part of mitosis, as chromosomes return to their semi-dispersed state, the tips of the 10 chromosomes coalesce and the nucleolus reforms (**Figure 6–45** and **Figure 6–46**). The transcription of the rRNA genes by RNA polymerase I is necessary for this process. As might be expected, the size of the nucleolus reflects the number of ribosomes that the cell is producing. Its size therefore varies greatly in different cells and can change in a single cell, occupying 25% of the total nuclear volume in cells that are making unusually large amounts of protein.

Ribosome assembly is a complex process, the most important features of which are outlined in **Figure 6–47**. In addition to its important role in ribosome biogenesis, the nucleolus is also the site where other RNAs are produced and other RNA–protein complexes are assembled. For example, the U6 snRNP, which functions in pre-mRNA splicing (see Figure 6–29), is composed of one RNA molecule and at least seven proteins. The U6 snRNA is chemically modified by snoRNAs in the nucleolus before its final assembly there into the U6 snRNP. Other important RNA–protein complexes, including telomerase (encountered in Chapter 5) and the signal recognition particle (which we discuss in Chapter 12), are also believed to be assembled at the nucleolus. Finally, the tRNAs (transfer RNAs) that carry the amino acids for protein synthesis are processed there as well; like the rRNA genes, those encoding tRNAs are clustered in the nucleolus. Thus, the nucleolus can be thought of as a large factory at which many different noncoding RNAs are transcribed, processed, and assembled with proteins to form a large variety of ribonucleoprotein complexes.

The Nucleus Contains a Variety of Subnuclear Structures

Although the nucleolus is the most prominent structure in the nucleus, several other nuclear bodies have been observed and studied (**Figure 6–48**). These include Cajal bodies (named for the scientist who first described them in 1906), GEMS (Gemini of Cajal bodies), and interchromatin granule clusters (also called

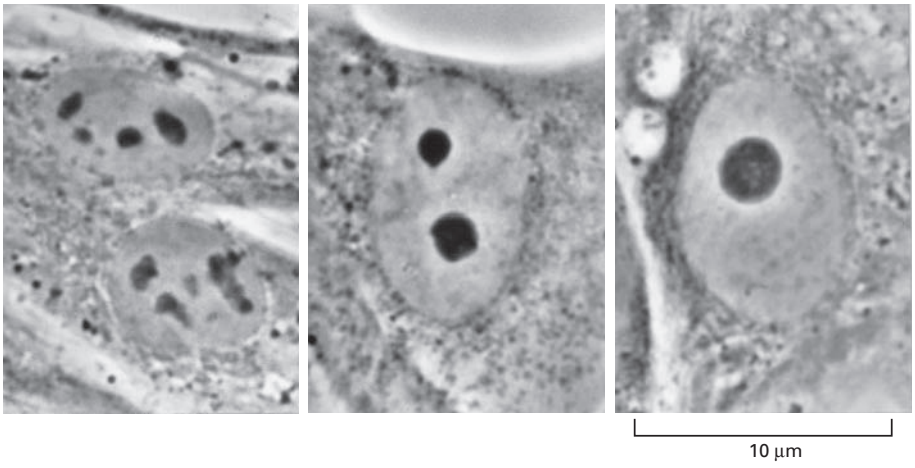


Figure 6–46 Nucleolar fusion. These light micrographs of human fibroblasts grown in culture show various stages of nucleolar fusion. After mitosis, each of the 10 human chromosomes that carry a cluster of rRNA genes begins to form a tiny nucleolus, but these rapidly coalesce as they grow to form the single large nucleolus typical of many interphase cells. (Courtesy of E.G. Jordan and J. McGovern.)

"speckles"). Like the nucleolus, these other nuclear structures lack membranes and are highly dynamic; their appearance is probably the result of the tight association of protein and RNA components involved in the synthesis, assembly, and storage of macromolecules involved in gene expression. Cajal bodies and GEMS resemble one another and are frequently paired in the nucleus; it is not clear whether they truly represent distinct structures. These are likely to be the locations in which snoRNAs and snRNAs undergo covalent modifications and final assembly with proteins. A group of guide RNAs, termed *small Cajal RNAs (scaRNAs)*, selects the sites of these modifications through base pairing. Cajal bodies/GEMS may also be sites where the snRNPs are recycled and their RNAs are "reset" after the rearrangements that occur during splicing (see p. 352). In contrast, the interchromatin granule clusters have been proposed to be stockpiles of fully mature snRNPs and other RNA processing components that are ready to be used in the production of mRNA (**Figure 6–49**).

Scientists have had difficulties in working out the function of these small subnuclear structures, in part because their appearances differ between organisms and can change dramatically as cells traverse the cell cycle or respond to changes in their environment. Much of the progress now being made depends on genetic tools—examination of the effects of designed mutations in model

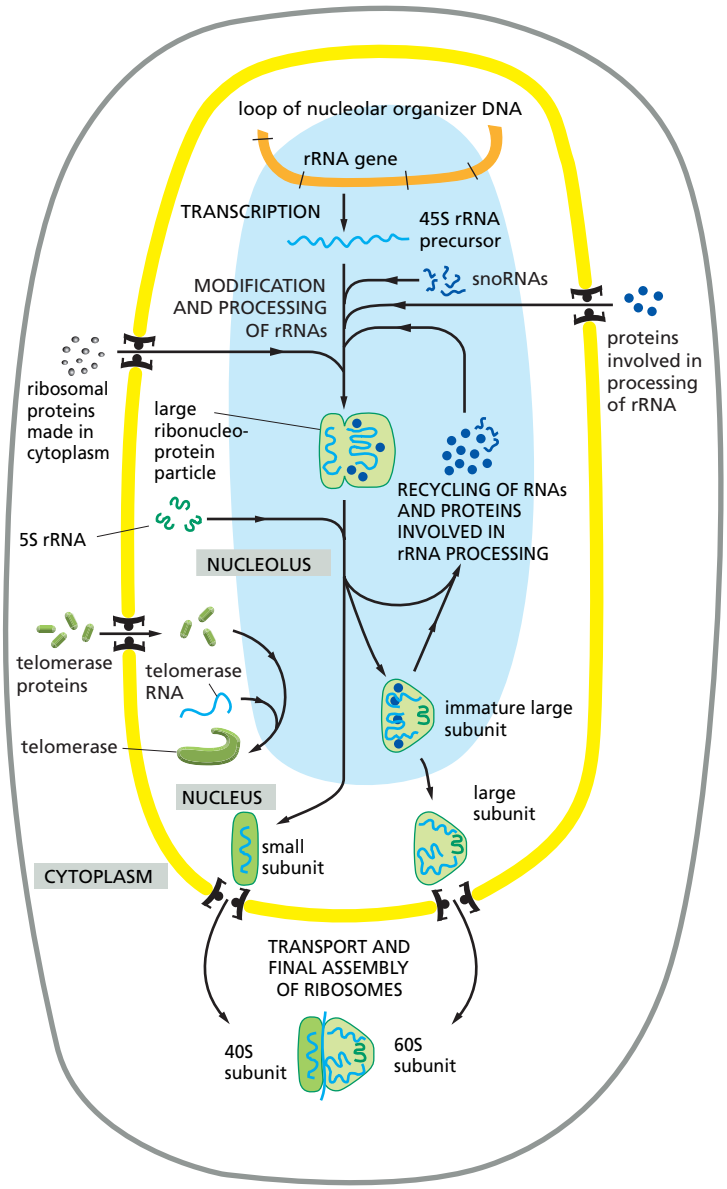


Figure 6–47 The function of the nucleolus in ribosome and other ribonucleoprotein synthesis. The 45S precursor rRNA is packaged in a large ribonucleoprotein particle containing many ribosomal proteins imported from the cytoplasm. While this particle remains at the nucleolus, selected pieces are added and others discarded as it is processed into immature large and small ribosomal subunits. The two ribosomal subunits are thought to attain their final functional form only as each is individually transported through the nuclear pores into the cytoplasm. Other ribonucleoprotein complexes, including telomerase shown here, are also assembled in the nucleolus.

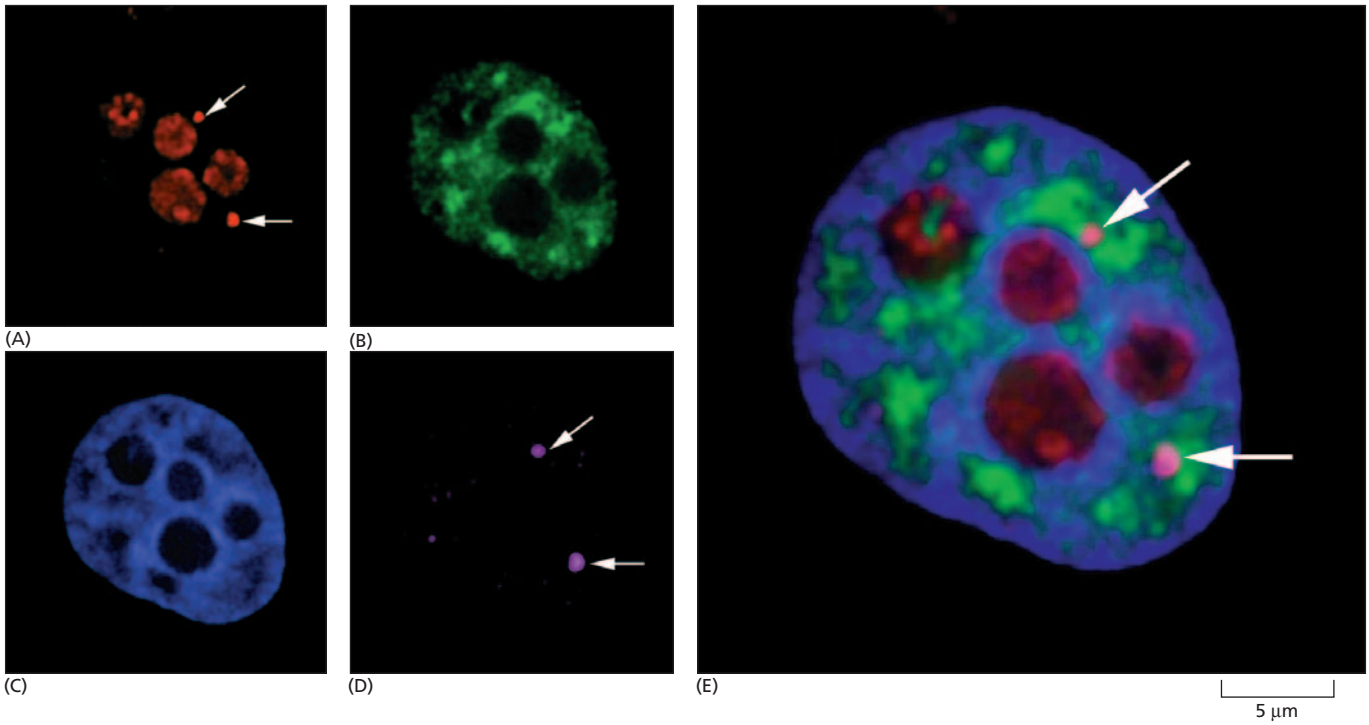


Figure 6–48 Visualization of some prominent nuclear bodies. (A)–(D) Micrographs of the same human cell nucleus, each processed to show a particular set of nuclear structures. (E) All four images enlarged and superimposed. (A) shows the location of the protein fibrillarin (a component of several snoRNPs), which is present at both nucleoli and Cajal bodies, the latter indicated by arrows. (B) shows interchromatin granule clusters or "speckles" detected by using antibodies against a protein involved in pre-mRNA splicing. (C) is stained to show bulk chromatin. (D) shows the location of the protein coilin, which is present at Cajal bodies (*arrows*; see also Figure 4–67). (From J.R. Swedlow and A.I. Lamond, *Gen. Biol.* 2:1–7, 2001. With permission from BioMed Central. Micrographs courtesy of Judith Sleeman.)

organisms or of spontaneous mutations in humans. As one example, GEMS contain the SMN (survival of motor neurons) protein. Certain mutations of the gene encoding this protein are the cause of inherited spinal muscular atrophy, a human disease characterized by a wasting away of the muscles. The disease seems to be caused by a defect in snRNP production. A more complete loss of snRNPs would be expected to be lethal.

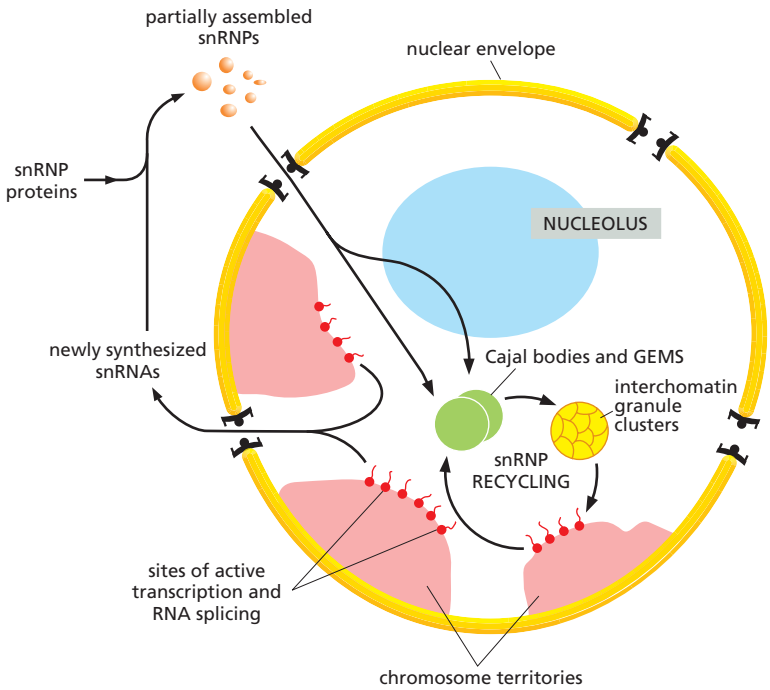


Figure 6–49 Schematic view of subnuclear structures. A typical vertebrate nucleus has several Cajal bodies, which are proposed to be the sites where snRNPs and snoRNPs undergo their final modifications. Interchromatin granule clusters are proposed to be storage sites for fully mature snRNPs. A typical vertebrate nucleus has 20–50 interchromatin granule clusters.

After their initial synthesis, snRNAs are exported from the nucleus to undergo 5′ and 3′ end-processing and assemble with the seven common snRNP proteins (called Sm proteins). These complexes are reimported into the nucleus and the snRNPs undergo their final modification by scaRNAs at Cajal bodies. In addition, snoRNAs chemically modify the U6 snRNP at the nucleolus. The sites of active transcription and splicing (approximately 2000–3000 sites per vertebrate nucleus) correspond to the "perichromatin fibers" seen under the electron microscope. (Adapted from J.D. Lewis and D. Tollervey, *Science* 288:1385–1389, 2000. With permission from AAAS.)

Given the importance of nuclear subdomains in RNA processing, it might have been expected that pre-mRNA splicing would occur in a particular location in the nucleus, as it requires numerous RNA and protein components. However, the assembly of splicing components on pre-mRNA is co-transcriptional; thus, splicing must occur at many locations along chromosomes. Although a typical mammalian cell may be expressing on the order of 15,000 genes, transcription and RNA splicing may be localized to only several thousand sites in the nucleus. These sites themselves are highly dynamic and probably result from the association of transcription and splicing components to create small "assembly lines" with a high local concentration of these components. Interchromatin granule clusters—which contain stockpiles of RNA-processing components—are often observed next to sites of transcription, as though poised to replenish supplies. Thus, the nucleus seems to be highly organized into subdomains, with snRNPs, snoRNPs, and other nuclear components moving between them in an orderly fashion according to the needs of the cell (see Figure 6–48; also see Figure 4–69).

Summary

Before the synthesis of a particular protein can begin, the corresponding mRNA molecule must be produced by transcription. Bacteria contain a single type of RNA polymerase (the enzyme that carries out the transcription of DNA into RNA). An mRNA molecule is produced when this enzyme initiates transcription at a promoter, synthesizes the RNA by chain elongation, stops transcription at a terminator, and releases both the DNA template and the completed mRNA molecule. In eucaryotic cells, the process of transcription is much more complex, and there are three RNA polymerases—polymerase I, II, and III—that are related evolutionarily to one another and to the bacterial polymerase.

RNA polymerase II synthesizes eucaryotic mRNA. This enzyme requires a series of additional proteins, the general transcription factors, to initiate transcription on a purified DNA template, and still more proteins (including chromatin-remodeling complexes and histone-modifying enzymes) to initiate transcription on its chromatin templates inside the cell.

During the elongation phase of transcription, the nascent RNA undergoes three types of processing events: a special nucleotide is added to its 5′ end (capping), intron sequences are removed from the middle of the RNA molecule (splicing), and the 3′ end of the RNA is generated (cleavage and polyadenylation). Each of these processes is initiated by proteins that travel along with RNA polymerase II by binding to sites on its long, extended C-terminal tail. Splicing is unusual in that many of its key steps are carried out by specialized RNA molecules rather than proteins. Properly processed mRNAs are passed through nuclear pore complexes into the cytosol, where they are translated into protein.

For some genes, RNA is the final product. In eucaryotes, these genes are usually transcribed by either RNA polymerase I or RNA polymerase III. RNA polymerase I makes the ribosomal RNAs. After their synthesis as a large precursor, the rRNAs are chemically modified, cleaved, and assembled into the two ribosomal subunits in the nucleolus—a distinct subnuclear structure that also helps to process some smaller RNA–protein complexes in the cell. Additional subnuclear structures (including Cajal bodies and interchromatin granule clusters) are sites where components involved in RNA processing are assembled, stored, and recycled.

FROM RNA TO PROTEIN

In the preceding section we have seen that the final product of some genes is an RNA molecule itself, such as those present in the snRNPs and in ribosomes. However, most genes in a cell produce mRNA molecules that serve as intermediaries on the pathway to proteins. In this section we examine how the cell converts the information carried in an mRNA molecule into a protein molecule. This feat of translation was a focus of attention of biologists in the late 1950s, when it

was posed as the "coding problem": how is the information in a linear sequence of nucleotides in RNA translated into the linear sequence of a chemically quite different set of units—the amino acids in proteins? This fascinating question stimulated great excitement among scientists at the time. Here was a cryptogram set up by nature that, after more than 3 billion years of evolution, could finally be solved by one of the products of evolution—human beings. And indeed, not only has the code been cracked step by step, but in the year 2000 the structure of the elaborate machinery by which cells read this code—the ribosome—was finally revealed in atomic detail.

An mRNA Sequence Is Decoded in Sets of Three Nucleotides

Once an mRNA has been produced by transcription and processing, the information present in its nucleotide sequence is used to synthesize a protein. Transcription is simple to understand as a means of information transfer: since DNA and RNA are chemically and structurally similar, the DNA can act as a direct template for the synthesis of RNA by complementary base-pairing. As the term *transcription* signifies, it is as if a message written out by hand is being converted, say, into a typewritten text. The language itself and the form of the message do not change, and the symbols used are closely related.

In contrast, the conversion of the information in RNA into protein represents a **translation** of the information into another language that uses quite different symbols. Moreover, since there are only 4 different nucleotides in mRNA and 20 different types of amino acids in a protein, this translation cannot be accounted for by a direct one-to-one correspondence between a nucleotide in RNA and an amino acid in protein. The nucleotide sequence of a gene, through the intermediary of mRNA, is translated into the amino acid sequence of a protein by rules that are known as the **genetic code**. This code was deciphered in the early 1960s.

The sequence of nucleotides in the mRNA molecule is read in consecutive groups of three. RNA is a linear polymer of four different nucleotides, so there are $4 \times 4 \times 4 = 64$ possible combinations of three nucleotides: the triplets AAA, AUA, AUG, and so on. However, only 20 different amino acids are commonly found in proteins. Either some nucleotide triplets are never used, or the code is redundant and some amino acids are specified by more than one triplet. The second possibility is, in fact, the correct one, as shown by the completely deciphered genetic code in **Figure 6–50**. Each group of three consecutive nucleotides in RNA is called a **codon**, and each codon specifies either one amino acid or a stop to the translation process.

This genetic code is used universally in all present-day organisms. Although a few slight differences in the code have been found, these are chiefly in the DNA of mitochondria. Mitochondria have their own transcription and protein synthesis systems that operate quite independently from those of the rest of the cell, and it is understandable that their small genomes have been able to accommodate minor changes to the code (discussed in Chapter 14).

	AGA									UUA					AGC				
	AGG									UUG					AGU				
GCA	CGA						GGA			CUA				CCA	UCA	ACA			GUA	
GCC	CGC						GGC		AUA	CUC				CCC	UCC	ACC			GUC	UAA
GCG	CGG	GAC	AAC	UGC	GAA	CAA	GGG	CAC	AUC	CUG	AAA		UUC	CCG	UCG	ACG		UAC	GUG	UAG
GCU	CGU	GAU	AAU	UGU	GAG	CAG	GGU	CAU	AUU	CUU	AAG	AUG	UUU	CCU	UCU	ACU	UGG	UAU	GUU	UGA
Ala	Arg	Asp	Asn	Cys	Glu	Gln	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val	stop
A	R	D	N	C	E	Q	G	H	I	L	K	M	F	P	S	T	W	Y	V	

Figure 6–50 The genetic code. The standard one-letter abbreviation for each amino acid is presented below its three-letter abbreviation (see Panel 3–1, pp. 128–129, for the full name of each amino acid and its structure). By convention, codons are always written with the 5′-terminal nucleotide to the left. Note that most amino acids are represented by more than one codon, and that there are some regularities in the set of codons that specifies each amino acid. Codons for the same amino acid tend to contain the same nucleotides at the first and second positions, and vary at the third position. Three codons do not specify any amino acid but act as termination sites (stop codons), signaling the end of the protein-coding sequence. One codon—AUG—acts both as an initiation codon, signaling the start of a protein-coding message, and also as the codon that specifies methionine.

In principle, an RNA sequence can be translated in any one of three different **reading frames**, depending on where the decoding process begins (**Figure 6–51**). However, only one of the three possible reading frames in an mRNA encodes the required protein. We see later how a special punctuation signal at the beginning of each RNA message sets the correct reading frame at the start of protein synthesis.

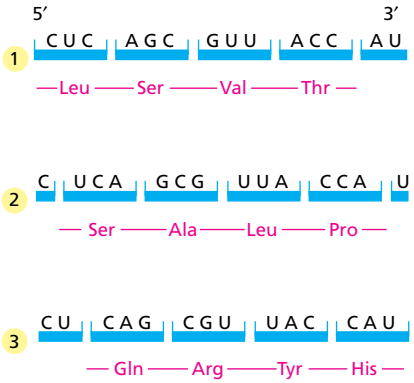


Figure 6–51 The three possible reading frames in protein synthesis. In the process of translating a nucleotide sequence *(blue)* into an amino acid sequence *(red),* the sequence of nucleotides in an mRNA molecule is read from the 5′ end to the 3′ end in consecutive sets of three nucleotides. In principle, therefore, the same RNA sequence can specify three completely different amino acid sequences, depending on the reading frame. In reality, however, only one of these reading frames contains the actual message.

tRNA Molecules Match Amino Acids to Codons in mRNA

The codons in an mRNA molecule do not directly recognize the amino acids they specify: the group of three nucleotides does not, for example, bind directly to the amino acid. Rather, the translation of mRNA into protein depends on adaptor molecules that can recognize and bind both to the codon and, at another site on their surface, to the amino acid. These adaptors consist of a set of small RNA molecules known as **transfer RNAs** (**tRNAs**), each about 80 nucleotides in length.

We saw earlier in this chapter that RNA molecules can fold into precise three-dimensional structures, and the tRNA molecules provide a striking example. Four short segments of the folded tRNA are double-helical, producing a molecule that looks like a cloverleaf when drawn schematically (**Figure 6–52**). For example, a 5′-GCUC-3′ sequence in one part of a polynucleotide chain can form a relatively strong association with a 5′-GAGC-3′ sequence in another region of the same molecule. The cloverleaf undergoes further folding to form a compact L-shaped structure that is held together by additional hydrogen bonds between different regions of the molecule.

Two regions of unpaired nucleotides situated at either end of the L-shaped molecule are crucial to the function of tRNA in protein synthesis. One of these

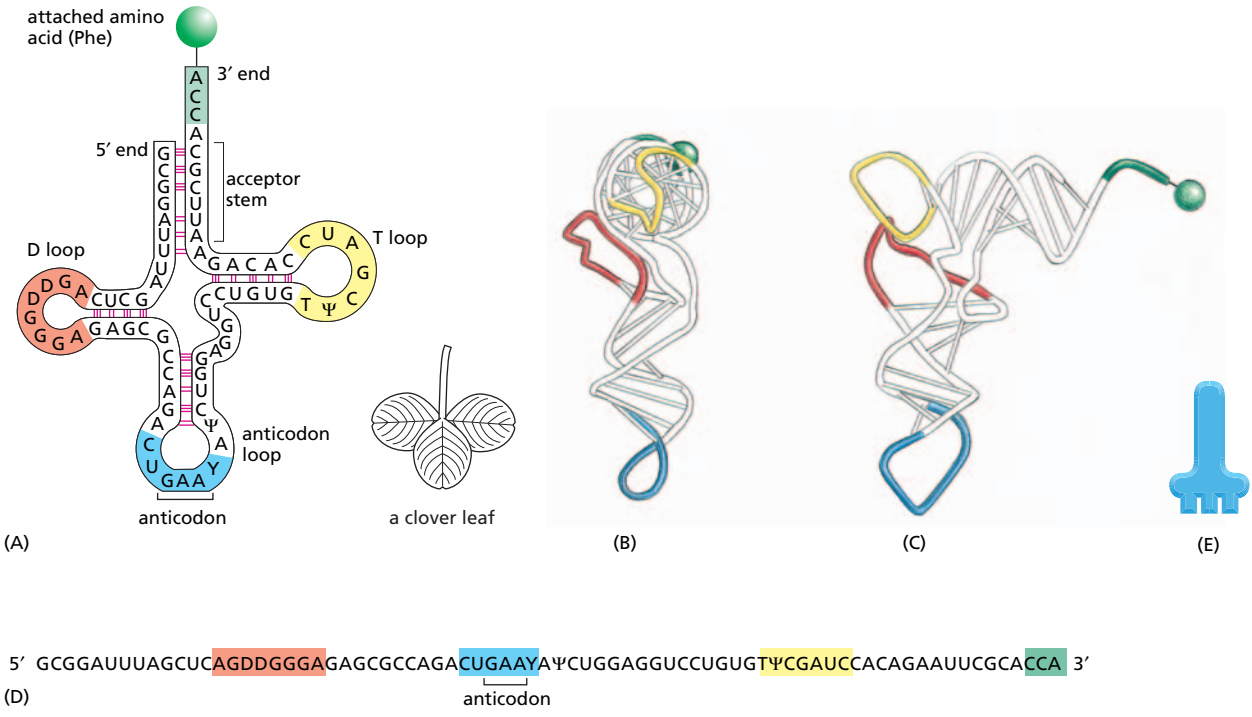


Figure 6–52 A tRNA molecule. A tRNA specific for the amino acid phenylalanine (Phe) is depicted in various ways. (A) The cloverleaf structure showing the complementary base-pairing *(red lines)* that creates the double-helical regions of the molecule. The anticodon is the sequence of three nucleotides that base-pairs with a codon in mRNA. The amino acid matching the codon/anticodon pair is attached at the 3′ end of the tRNA. tRNAs contain some unusual bases, which are produced by chemical modification after the tRNA has been synthesized. For example, the bases denoted ψ (pseudouridine—see Figure 6–43) and D (dihydrouridine—see Figure 6–55) are derived from uracil. (B and C) Views of the L-shaped molecule, based on x-ray diffraction analysis. Although this diagram shows the tRNA for the amino acid phenylalanine, all other tRNAs have similar structures. <CGCA> (D) The linear nucleotide sequence of the molecule, color-coded to match (A), (B), and (C). (E) The tRNA icon we sue in this book.

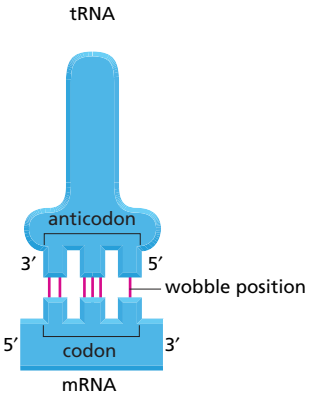
regions forms the **anticodon**, a set of three consecutive nucleotides that pairs with the complementary codon in an mRNA molecule. The other is a short single-stranded region at the 3′ end of the molecule; this is the site where the amino acid that matches the codon is attached to the tRNA.

We have seen in the previous section that the genetic code is redundant; that is, several different codons can specify a single amino acid (see Figure 6–50). This redundancy implies either that there is more than one tRNA for many of the amino acids or that some tRNA molecules can base-pair with more than one codon. In fact, both situations occur. Some amino acids have more than one tRNA and some tRNAs are constructed so that they require accurate base-pairing only at the first two positions of the codon and can tolerate a mismatch (or *wobble*) at the third position (**Figure 6–53**). This wobble base-pairing explains why so many of the alternative codons for an amino acid differ only in their third nucleotide (see Figure 6–50). In bacteria, wobble base-pairings make it possible to fit the 20 amino acids to their 61 codons with as few as 31 kinds of tRNA molecules. The exact number of different kinds of tRNAs, however, differs from one species to the next. For example, humans have nearly 500 tRNA genes but, among them, only 48 different anticodons are represented.

tRNAs Are Covalently Modified Before They Exit from the Nucleus

Like most other eucaryotic RNAs, tRNAs are covalently modified before they are allowed to exit from the nucleus. Eucaryotic tRNAs are synthesized by RNA polymerase III. Both bacterial and eucaryotic tRNAs are typically synthesized as larger precursor tRNAs, which are then trimmed to produce the mature tRNA. In addition, some tRNA precursors (from both bacteria and eucaryotes) contain introns that must be spliced out. This splicing reaction differs chemically from pre-mRNA splicing; rather than generating a lariat intermediate, tRNA splicing uses a cut-and-paste mechanism that is catalyzed by proteins (**Figure 6–54**). Trimming and splicing both require the precursor tRNA to be correctly folded in its cloverleaf configuration. Because misfolded tRNA precursors will not be processed properly, the trimming and splicing reactions are thought to act as quality-control steps in the generation of tRNAs.

All tRNAs are modified chemically—nearly 1 in 10 nucleotides in each mature tRNA molecule is an altered version of a standard G, U, C, or A ribonucleotide. Over 50 different types of tRNA modifications are known; a few are shown in **Figure 6–55**. Some of the modified nucleotides—most notably inosine, produced by the deamination of adenosine—affect the conformation and base-pairing of the anticodon and thereby facilitate the recognition of the appropriate mRNA codon by the tRNA molecule (see Figure 6–53). Others affect the accuracy with which the tRNA is attached to the correct amino acid.



bacteria

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

eucaryotes

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

Figure 6–53 Wobble base-pairing between codons and anticodons. If the nucleotide listed in the first column is present at the third, or wobble, position of the codon, it can base-pair with any of the nucleotides listed in the second column. Thus, for example, when inosine (I) is present in the wobble position of the tRNA anticodon, the tRNA can recognize any one of three different codons in bacteria and either of two codons in eucaryotes. The inosine in tRNAs is formed from the deamination of guanine (see Figure 6–55), a chemical modification that takes place after the tRNA has been synthesized. The nonstandard base pairs, including those made with inosine, are generally weaker than conventional base pairs. Note that codon–anticodon base pairing is more stringent at positions 1 and 2 of the codon: here only conventional base pairs are permitted. The differences in wobble base-pairing interactions between bacteria and eucaryotes presumably result from subtle structural differences between bacterial and eucaryotic ribosomes, the molecular machines that perform protein synthesis. (Adapted from C. Guthrie and J. Abelson, in The Molecular Biology of the Yeast *Saccharomyces*: Metabolism and Gene Expression, pp. 487–528. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press, 1982.)

Figure 6–54 Structure of a tRNA-splicing endonuclease docked to a precursor tRNA. The endonuclease (a four-subunit enzyme) removes the tRNA intron *(blue)*. A second enzyme, a multifunctional tRNA ligase (not shown), then joins the two tRNA halves together. (Courtesy of Hong Li, Christopher Trotta and John Abelson.)

Specific Enzymes Couple Each Amino Acid to Its Appropriate tRNA Molecule

We have seen that, to read the genetic code in DNA, cells make a series of different tRNAs. We now consider how each tRNA molecule becomes linked to the one amino acid in 20 that is its appropriate partner. Recognition and attachment of the correct amino acid depends on enzymes called **aminoacyl-tRNA synthetases**, which covalently couple each amino acid to its appropriate set of tRNA molecules (**Figure 6–56** and **Figure 6–57**). Most cells have a different synthetase enzyme for each amino acid (that is, 20 synthetases in all); one attaches glycine to all tRNAs that recognize codons for glycine, another attaches alanine to all tRNAs that recognize codons for alanine, and so on. Many bacteria, however, have fewer than 20 synthetases, and the same synthetase enzyme is responsible for coupling more than one amino acid to the appropriate tRNAs. In these cases, a single synthetase places the identical amino acid on two different types of tRNAs, only one of which has an anticodon that matches the amino acid. A second enzyme then chemically modifies each "incorrectly" attached amino acid so that it now corresponds to the anticodon displayed by its covalently linked tRNA.

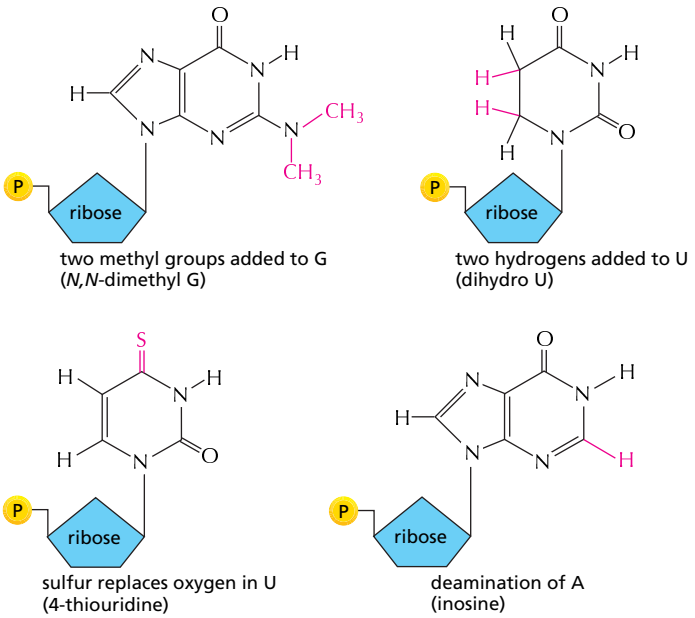


Figure 6–55 A few of the unusual nucleotides found in tRNA molecules. These nucleotides are produced by covalent modification of a normal nucleotide after it has been incorporated into an RNA chain. Two other types of modified nucleotides are shown in Figure 6–43. In most tRNA molecules about 10% of the nucleotides are modified (see Figure 6–52).

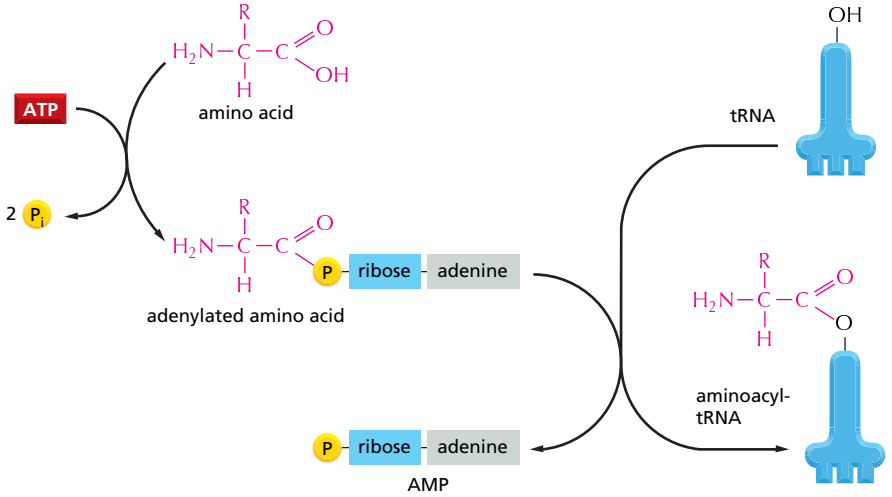


Figure 6–56 Amino acid activation. An amino acid is activated for protein synthesis by an aminoacyl-tRNA synthetase enzyme in two steps. As indicated, the energy of ATP hydrolysis is used to attach each amino acid to its tRNA molecule in a high-energy linkage. The amino acid is first activated through the linkage of its carboxyl group directly to an AMP moiety, forming an *adenylated amino acid*; the linkage of the AMP, normally an unfavorable reaction, is driven by the hydrolysis of the ATP molecule that donates the AMP. Without leaving the synthetase enzyme, the AMP-linked carboxyl group on the amino acid is then transferred to a hydroxyl group on the sugar at the 3′ end of the tRNA molecule. This transfer joins the amino acid by an activated ester linkage to the tRNA and forms the final aminoacyl-tRNA molecule. The synthetase enzyme is not shown in this diagram.

The synthetase-catalyzed reaction that attaches the amino acid to the 3′ end of the tRNA is one of many reactions coupled to the energy-releasing hydrolysis of ATP (see pp. 79–81), and it produces a high-energy bond between the tRNA and the amino acid. The energy of this bond is used at a later stage in protein synthesis to link the amino acid covalently to the growing polypeptide chain.

The aminoacyl-tRNA synthetase enzymes and the tRNAs are equally important in the decoding process (**Figure 6–58**). This was established by an experiment in which one amino acid (cysteine) was chemically converted into a different amino acid (alanine) after it already had been attached to its specific tRNA. When such "hybrid" aminoacyl-tRNA molecules were used for protein synthesis in a cell-free system, the wrong amino acid was inserted at every point in the protein chain where that tRNA was used. Although, as we shall see, cells have several quality control mechanisms to avoid this type of mishap, the experiment establishes that the genetic code is translated by two sets of adaptors that act sequentially. Each matches one molecular surface to another with great specificity, and it is their combined action that associates each sequence of three nucleotides in the mRNA molecule—that is, each codon—with its particular amino acid.

Editing by tRNA Synthetases Ensures Accuracy

Several mechanisms working together ensure that the tRNA synthetase links the correct amino acid to each tRNA. The synthetase must first select the correct amino acid, and most synthetases do so by a two-step mechanism. First, the correct amino acid has the highest affinity for the active-site pocket of its synthetase

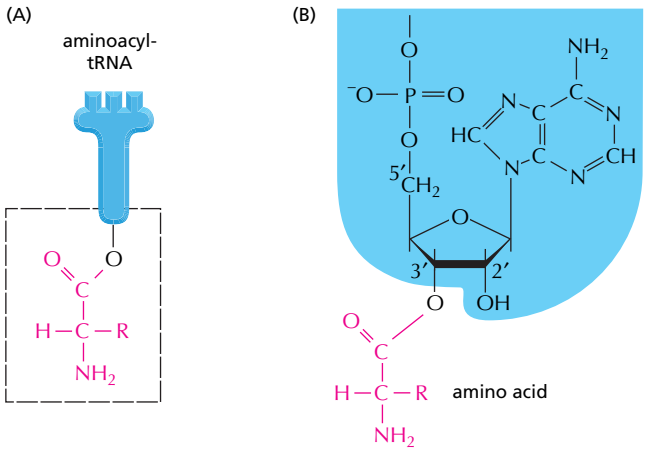
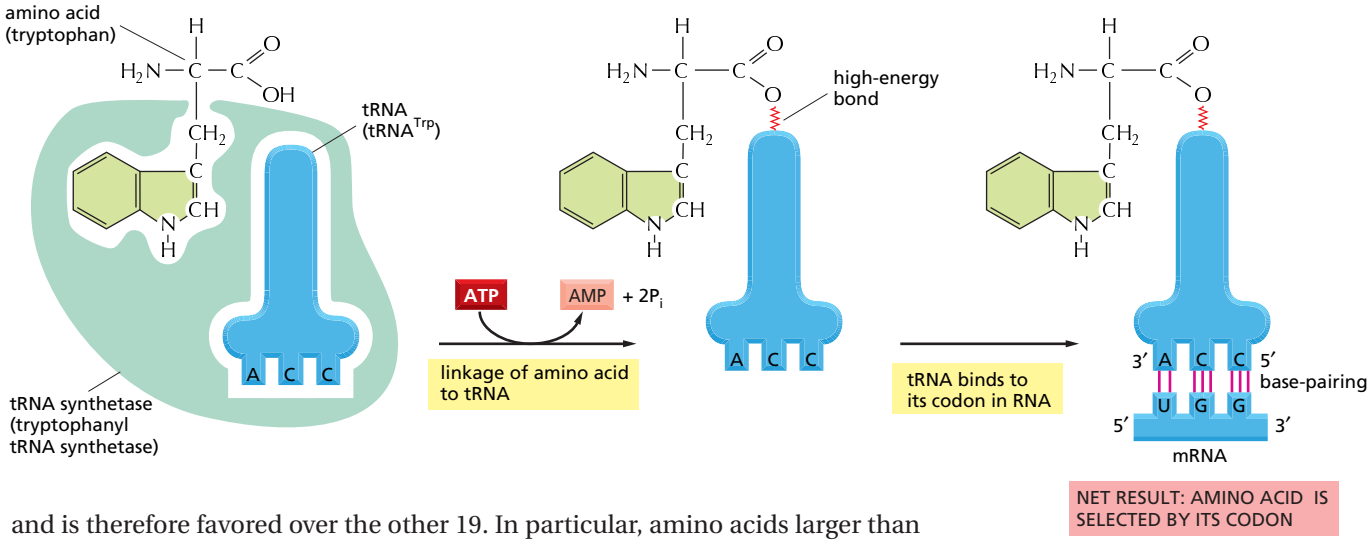


Figure 6–57 The structure of the aminoacyl-tRNA linkage. The carboxyl end of the amino acid forms an ester bond to ribose. Because the hydrolysis of this ester bond is associated with a large favorable change in free energy, an amino acid held in this way is said to be activated. (A) Schematic drawing of the structure. The amino acid is linked to the nucleotide at the 3′ end of the tRNA (see Figure 6–52). (B) Actual structure corresponding to the boxed region in (A). There are two major classes of synthetase enzymes: one links the amino acid directly to the 3′-OH group of the ribose, and the other links it initially to the 2′-OH group. In the latter case, a subsequent transesterification reaction shifts the amino acid to the 3′ position. As in Figure 6–56, the "R group" indicates the side chain of the amino acid.



and is therefore favored over the other 19. In particular, amino acids larger than the correct one are effectively excluded from the active site. However, accurate discrimination between two similar amino acids, such as isoleucine and valine (which differ by only a methyl group), is very difficult to achieve by a one-step recognition mechanism. A second discrimination step occurs after the amino acid has been covalently linked to AMP (see Figure 6–56). When tRNA binds the synthetase, it tries to force the amino acid into a second pocket in the synthetase, the precise dimensions of which exclude the correct amino acid but allow access by closely related amino acids. Once an amino acid enters this editing pocket, it is hydrolyzed from the AMP (or from the tRNA itself if the aminoacyl-tRNA bond has already formed), and is released from the enzyme. This hydrolytic editing, which is analogous to the exonucleolytic proofreading by DNA polymerases (**Figure 6–59**), raises the overall accuracy of tRNA charging to approximately one mistake in 40,000 couplings.

Figure 6–58 The genetic code is translated by means of two adaptors that act one after another. The first adaptor is the aminoacyl-tRNA synthetase, which couples a particular amino acid to its corresponding tRNA; the second adaptor is the tRNA molecule itself, whose *anticodon* forms base pairs with the appropriate *codon* on the mRNA. An error in either step would cause the wrong amino acid to be incorporated into a protein chain. In the sequence of events shown, the amino acid tryptophan (Trp) is selected by the codon UGG on the mRNA.

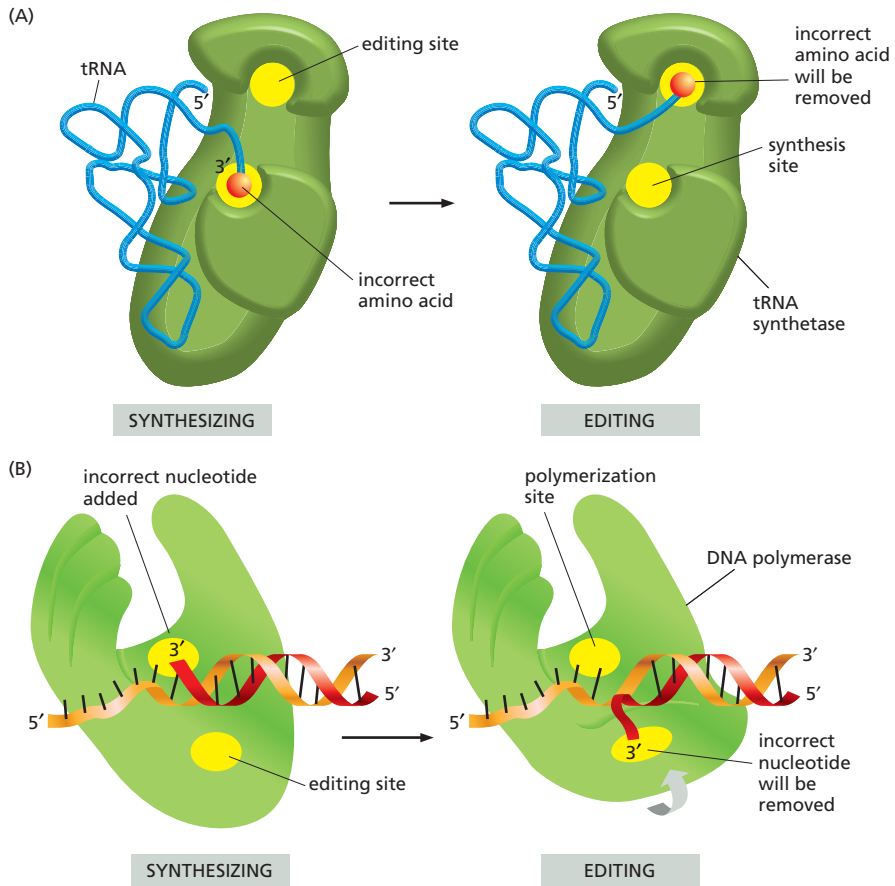


Figure 6–59 Hydrolytic editing. (A) tRNA synthetases remove their own coupling errors through hydrolytic editing of incorrectly attached amino acids. As described in the text, the correct amino acid is rejected by the editing site. (B) The error-correction process performed by DNA polymerase shows some similarities; however, it differs in so far as the removal process depends strongly on a mispairing with the template (see Figure 5–8).

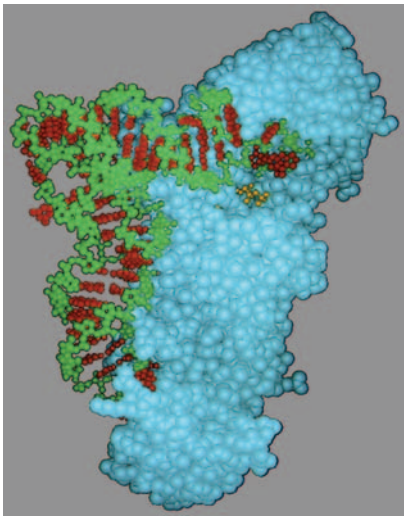


Figure 6–60 The recognition of a tRNA molecule by its aminoacyl-tRNA synthetase. For this tRNA ($tRNA^{Gln}$), specific nucleotides in both the anticodon *(bottom)* and the amino acid-accepting arm allow the correct tRNA to be recognized by the synthetase enzyme *(blue)*. A bound ATP molecule is *yellow*. (Courtesy of Tom Steitz.)

The tRNA synthetase must also recognize the correct set of tRNAs, and extensive structural and chemical complementarity between the synthetase and the tRNA allows the synthetase to probe various features of the tRNA (**Figure 6–60**). Most tRNA synthetases directly recognize the matching tRNA anticodon; these synthetases contain three adjacent nucleotide-binding pockets, each of which is complementary in shape and charge to a nucleotide in the anticodon. For other synthetases, the nucleotide sequence of the acceptor stem is the key recognition determinant. In most cases, however, the synthetase "reads" the nucleotides at several different positions on the tRNA.

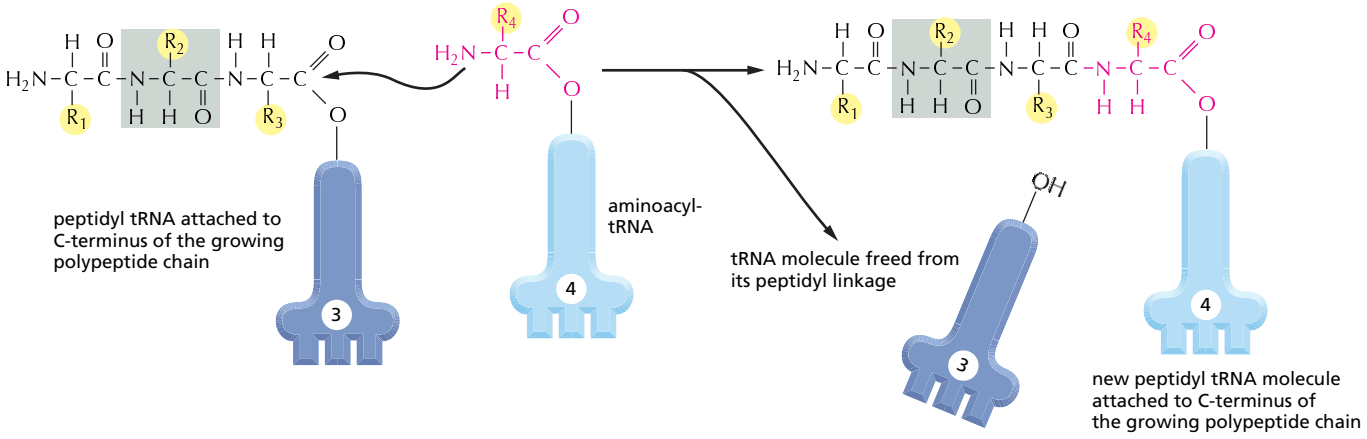
Amino Acids Are Added to the C-terminal End of a Growing Polypeptide Chain

Having seen that amino acids are first coupled to tRNA molecules, we now turn to the mechanism that joins amino acids together to form proteins. The fundamental reaction of protein synthesis is the formation of a peptide bond between the carboxyl group at the end of a growing polypeptide chain and a free amino group on an incoming amino acid. Consequently, a protein is synthesized stepwise from its N-terminal end to its C-terminal end. Throughout the entire process the growing carboxyl end of the polypeptide chain remains activated by its covalent attachment to a tRNA molecule (forming a peptidyl-tRNA). Each addition disrupts this high-energy covalent linkage, but immediately replaces it with an identical linkage on the most recently added amino acid (**Figure 6–61**). In this way, each amino acid added carries with it the activation energy for the addition of the next amino acid rather than the energy for its own addition—an example of the "head growth" type of polymerization described in Figure 2–68.

The RNA Message Is Decoded in Ribosomes

The synthesis of proteins is guided by information carried by mRNA molecules. To maintain the correct reading frame and to ensure accuracy (about 1 mistake every 10,000 amino acids), protein synthesis is performed in the **ribosome**, a complex catalytic machine made from more than 50 different proteins (the *ribosomal proteins*) and several RNA molecules, the **ribosomal RNAs** (**rRNAs**). <CGCC> A typical eucaryotic cell contains millions of ribosomes in its cytoplasm (**Figure 6–62**). Eucaryotic ribosome subunits are assembled at the nucleolus, when newly transcribed and modified rRNAs associate with ribosomal

Figure 6–61 The incorporation of an amino acid into a protein. A polypeptide chain grows by the stepwise addition of amino acids to its C-terminal end. The formation of each peptide bond is energetically favorable because the growing C-terminus has been activated by the covalent attachment of a tRNA molecule. The peptidyl-tRNA linkage that activates the growing end is regenerated during each addition. The amino acid side chains have been abbreviated as R_1, R_2, R_3, and R_4; as a reference point, all of the atoms in the second amino acid in the polypeptide chain are shaded *gray*. The figure shows the addition of the fourth amino acid *(red)* to the growing chain.



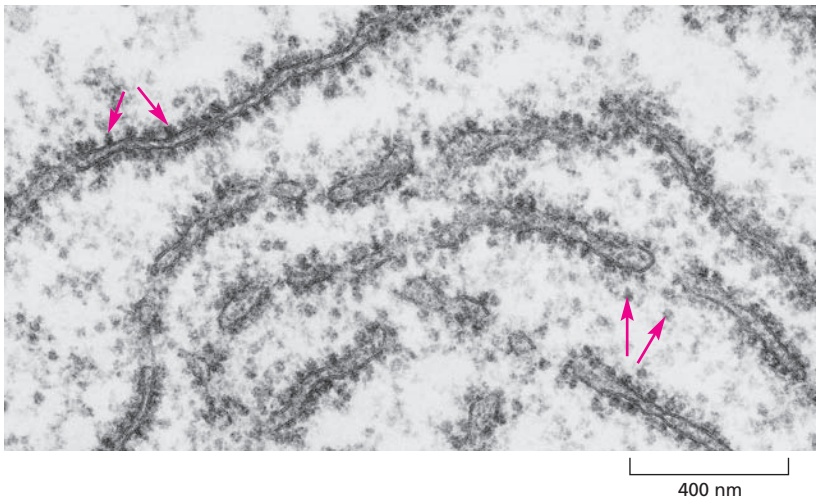


Figure 6–62 Ribosomes in the cytoplasm of a eucaryotic cell. This electron micrograph shows a thin section of a small region of cytoplasm. The ribosomes appear as black dots *(red arrows)*. Some are free in the cytosol; others are attached to membranes of the endoplasmic reticulum. (Courtesy of Daniel S. Friend.)

proteins, which have been transported into the nucleus after their synthesis in the cytoplasm. The two ribosomal subunits are then exported to the cytoplasm, where they join together to synthesize proteins.

Eucaryotic and procaryotic ribosomes have similar designs and functions. Both are composed of one large and one small subunit that fit together to form a complete ribosome with a mass of several million daltons (**Figure 6–63**). The small subunit provides the framework on which the tRNAs can be accurately

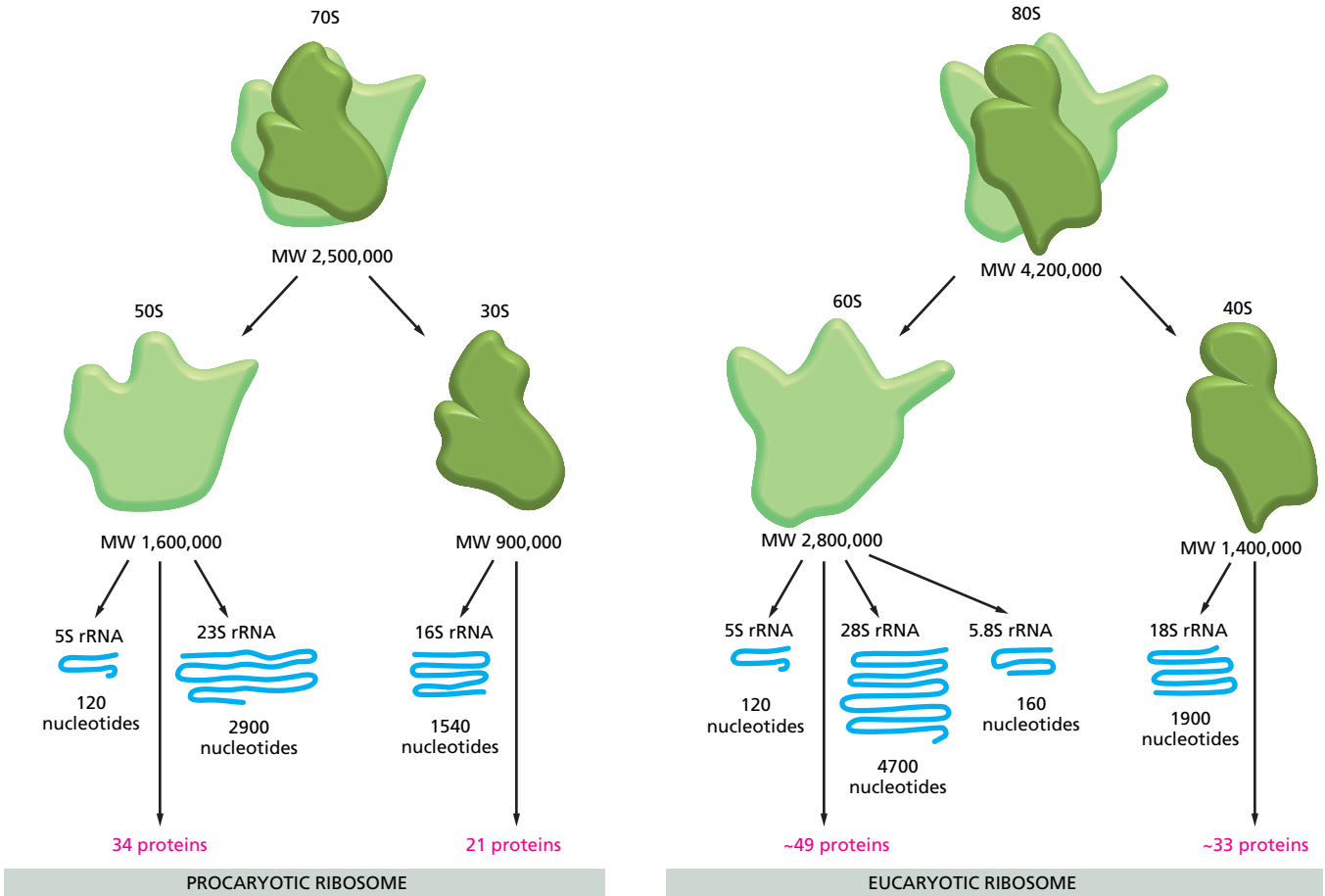


Figure 6–63 A comparison of procaryotic and eucaryotic ribosomes. Despite differences in the number and size of their rRNA and protein components, both procaryotic and eucaryotic ribosomes have nearly the same structure and they function similarly. Although the 18S and 28S rRNAs of the eucaryotic ribosome contain many nucleotides not present in their bacterial counterparts, these nucleotides are present as multiple insertions that form extra domains and leave the basic structure of each rRNA largely unchanged.

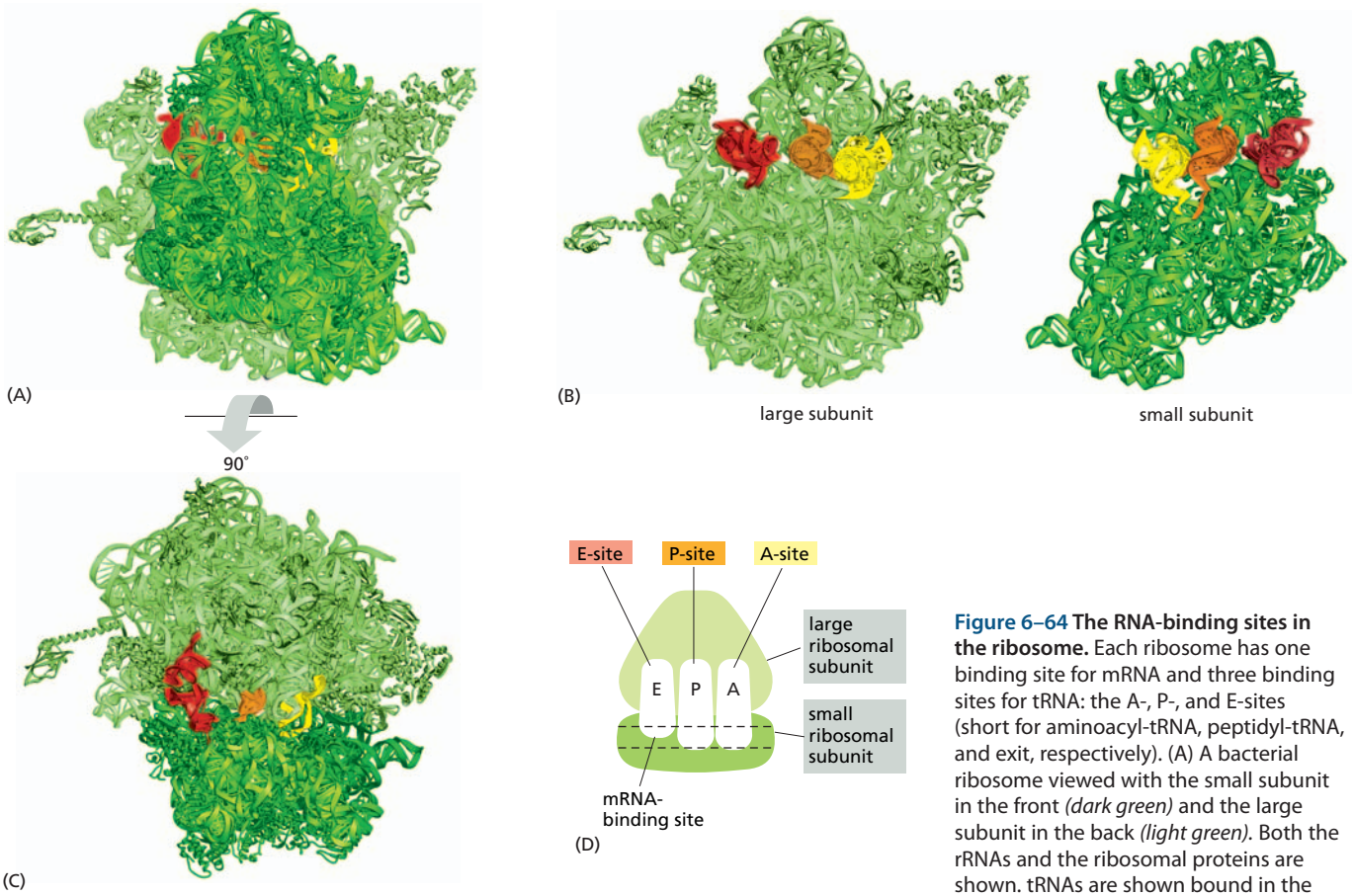


Figure 6–64 The RNA-binding sites in the ribosome. Each ribosome has one binding site for mRNA and three binding sites for tRNA: the A-, P-, and E-sites (short for aminoacyl-tRNA, peptidyl-tRNA, and exit, respectively). (A) A bacterial ribosome viewed with the small subunit in the front *(dark green)* and the large subunit in the back *(light green)*. Both the rRNAs and the ribosomal proteins are shown. tRNAs are shown bound in the E-site *(red)*, the P-site *(orange)* and the A-site *(yellow)*. Although all three tRNA sites are shown occupied here, during the process of protein synthesis not more than two of these sites are thought to contain tRNA molecules at any one time (see Figure 6–66). (B) Large and small ribosomal subunits arranged as though the ribosome in (A) were opened like a book. (C) The ribosome in (A) rotated through 90º and viewed with the large subunit on top and small subunit on the bottom. (D) Schematic representation of a ribosome (in the same orientation as C), which will be used in subsequent figures. (A, B, and C, adapted from M.M. Yusupov et al., *Science* 292:883–896, 2001. With permission from AAAS; courtesy of Albion Baucom and Harry Noller.)

matched to the codons of the mRNA (see Figure 6–58), while the large subunit catalyzes the formation of the peptide bonds that link the amino acids together into a polypeptide chain (see Figure 6–61).

When not actively synthesizing proteins, the two subunits of the ribosome are separate. They join together on an mRNA molecule, usually near its 5′ end, to initiate the synthesis of a protein. The mRNA is then pulled through the ribosome; as its codons enter the core of the ribosome, the mRNA nucleotide sequence is translated into an amino acid sequence using the tRNAs as adaptors to add each amino acid in the correct sequence to the end of the growing polypeptide chain. When a stop codon is encountered, the ribosome releases the finished protein, and its two subunits separate again. These subunits can then be used to start the synthesis of another protein on another mRNA molecule.

Ribosomes operate with remarkable efficiency: in one second, a single ribosome of a eucaryotic cell adds about 2 amino acids to a polypeptide chain; the ribosomes of bacterial cells operate even faster, at a rate of about 20 amino acids per second. How does the ribosome choreograph the many coordinated movements required for efficient translation? A ribosome contains four binding sites for RNA molecules: one is for the mRNA and three (called the A-site, the P-site, and the E-site) are for tRNAs (**Figure 6–64**). A tRNA molecule is held tightly at the A- and P-sites only if its anticodon forms base pairs with a complementary codon (allowing for wobble) on the mRNA molecule that is threaded through the ribosome (**Figure 6–65**). The A- and P-sites are close enough together for their two tRNA molecules to be forced to form base pairs with adjacent codons on the mRNA molecule. This feature of the ribosome maintains the correct reading frame on the mRNA.

Once protein synthesis has been initiated, each new amino acid is added to the elongating chain in a cycle of reactions containing four major steps: tRNA

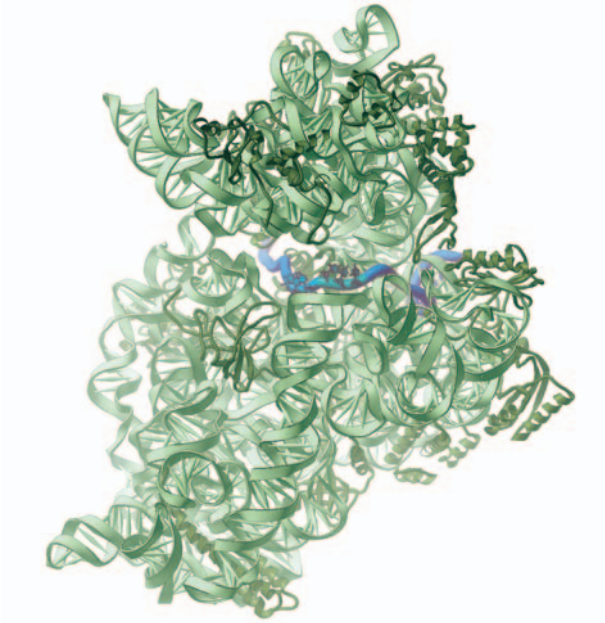


Figure 6–65 The path of mRNA *(blue)* through the small ribosomal subunit. The orientation is the same as that in the right-hand panel of Figure 6–64B. (Courtesy of Harry F. Noller, based on data in G.Z. Yusopova et al., *Cell* 106:233–241, 2001. With permission from Elsevier.)

binding, peptide bond formation, large subunit and small subunit translocation. As a result of the two translocation steps, the entire ribosome moves three nucleotides along the mRNA and is positioned to start the next cycle. (**Figure 6–66**). Our description of the chain elongation process begins at a point at which some amino acids have already been linked together and there is a tRNA molecule in the P-site on the ribosome, covalently joined to the end of the growing polypeptide. In step 1, a tRNA carrying the next amino acid in the chain binds to the ribosomal A-site by forming base pairs with the mRNA codon positioned there, so that the P-site and the A-site contain adjacent bound tRNAs. In step 2, the carboxyl end of the polypeptide chain is released from the tRNA at the P-site (by breakage of the high-energy bond between the tRNA and its amino acid) and joined to the free amino group of the amino acid linked to the tRNA at the A-site, forming a new peptide bond. This central reaction of protein synthesis is catalyzed by a *peptidyl transferase* contained in the large ribosomal subunit. In step 3, the large subunit moves relative to the mRNA held by the small subunit, thereby shifting the acceptor stems of the two tRNAs to the E- and P-sites of the large subunit. In step 4, another series of conformational changes moves the small subunit and its bound mRNA exactly three nucleotides, resetting the ribosome so it is ready to receive the next aminoacyl-tRNA. Step 1 is then repeated with a new incoming aminoacyl-tRNA, and so on. <CGTT>

This four-step cycle is repeated each time an amino acid is added to the polypeptide chain, as the chain grows from its amino to its carboxyl end.

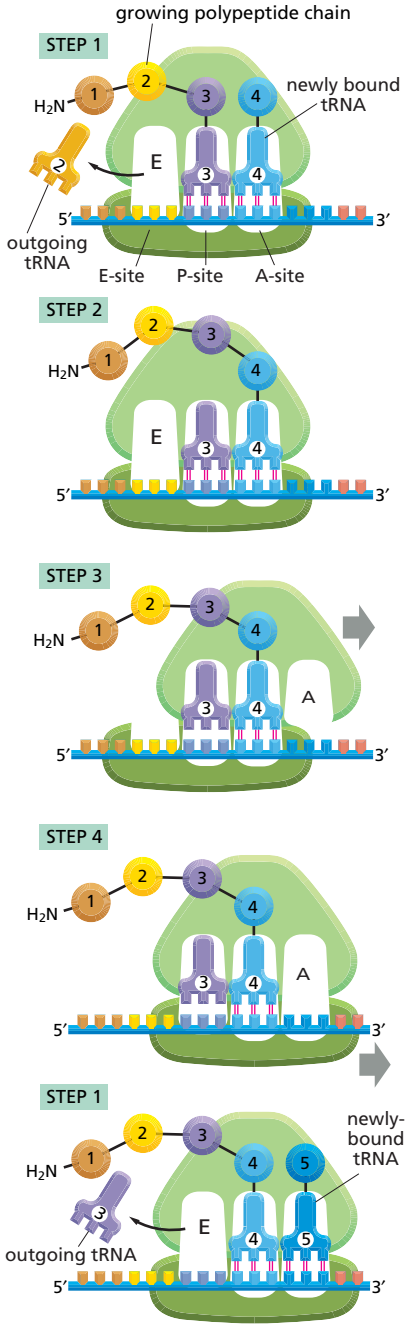


Figure 6–66 Translating an mRNA molecule. Each amino acid added to the growing end of a polypeptide chain is selected by complementary base-pairing between the anticodon on its attached tRNA molecule and the next codon on the mRNA chain. Because only one of the many types of tRNA molecules in a cell can base-pair with each codon, the codon determines the specific amino acid to be added to the growing polypeptide chain. The four-step cycle shown is repeated over and over during the synthesis of a protein. In step 1, an aminoacyl-tRNA molecule binds to a vacant A-site on the ribosome and a spent tRNA molecule dissociates from the E-site. In step 2, a new peptide bond is formed. In step 3, the large subunit translocates relative to the small subunit, leaving the two tRNAs in hybrid sites: P on the large subunit and A on the small, for one; E on the large subunit and P on the small, for the other. In step 4, the small subunit translocates carrying its mRNA a distance of three nucleotides through the ribosome. This "resets" the ribosome with a fully empty A-site, ready for the next aminoacyl-tRNA molecule to bind. As indicated, the mRNA is translated in the 5′-to-3′ direction, and the N-terminal end of a protein is made first, with each cycle adding one amino acid to the C-terminus of the polypeptide chain.

Elongation Factors Drive Translation Forward and Improve Its Accuracy

The basic cycle of polypeptide elongation shown in outline in Figure 6–66 has an additional feature that makes translation especially efficient and accurate. Two *elongation factors* enter and leave the ribosome during each cycle, each hydrolyzing GTP to GDP and undergoing conformational changes in the process. These factors are called EF-Tu and EF-G in bacteria, and EF1 and EF2 in eucaryotes. Under some conditions *in vitro,* ribosomes can be forced to synthesize proteins without the aid of these elongation factors and GTP hydrolysis, but this synthesis is very slow, inefficient, and inaccurate. Coupling the GTP hydrolysis-driven changes in the elongation factors to transitions between different states of the ribosome speeds up protein synthesis enormously. Although these ribosomal states are not yet understood in detail, they almost certainly involve RNA structure rearrangements in the ribosome core. The cycles of elongation factor association, GTP hydrolysis, and dissociation ensure that all such changes occur in the "forward" direction so that translation can proceed efficiently (**Figure 6–67**).

As shown previously, EF-Tu simultaneously binds GTP and aminoacyl-tRNAs (see Figure 3–74). In addition to helping move translation forward, EF-Tu (EF1 in eucaryotes) increases the accuracy of translation in several ways. First, as it escorts an incoming aminoacyl-tRNA to the ribosome, EF-Tu checks whether the tRNA–amino acid match is corrrect. Exactly how this is accomplished is not well understood. According to one idea, correct tRNA–amino acid matches have a narrowly defined affinity for EF-Tu, which allows EF-Tu to discriminate, albeit crudely, among many different amino acid–tRNA combinations, selectively bringing the correct ones with it into the ribosome. Second, EF-Tu monitors the initial interaction between the anticodon of an incoming aminoacyl-tRNA and the codon of the mRNA in the A-site. Aminoacyl-tRNAs are "bent" when bound to the GTP-form of EF-Tu; this bent conformation allows codon pairing but prevents incorporation of the amino acid into the growing polypeptide chain. However, if the codon–anticodon match is correct, the ribosome rapidly triggers the hydrolysis of the GTP molecule, whereupon EF-Tu releases its grip on the tRNA and dissociates from the ribosome, allowing the tRNA to donate its amino acid for protein synthesis. But how is the "correctness" of the codon–anticodon match assessed? This feat is carried out by the ribosome itself through an RNA-based mechanism. The rRNA in the small subunit of the ribosome forms a series of hydrogen bonds with the codon–anticodon pair that allows determination of its correctness (**Figure 6–68**). In essence, the rRNA folds around the codon–anticodon pair, and its final closure—which occurs only when the correct anticodon is in place—triggers GTP hydrolysis. Remarkably, this induced fit mechanism can distinguish correct from incorrect codon–anticodon interactions despite the rules for wobble base-pairing summarized in Figure 6–53. From this example, as for RNA splicing, one gets a sense of the highly sophisticated forms of molecular recognition that can be achieved solely by RNA.

The interactions of EF-Tu, tRNA, and the ribosome just described introduce critical proofreading steps into protein synthesis at the initial tRNA selection stage. But after GTP is hydrolyzed and EF-Tu dissociates from the ribosome, there is an additional opportunity for the ribosome to prevent an incorrect amino acid from being added to the growing chain. Following GTP hydrolysis, there is a short time delay as the amino acid carried by the tRNA moves into position on the ribosome. This time delay is shorter for correct than incorrect codon–anticodon pairs. Moreover, incorrectly matched tRNAs dissociate more

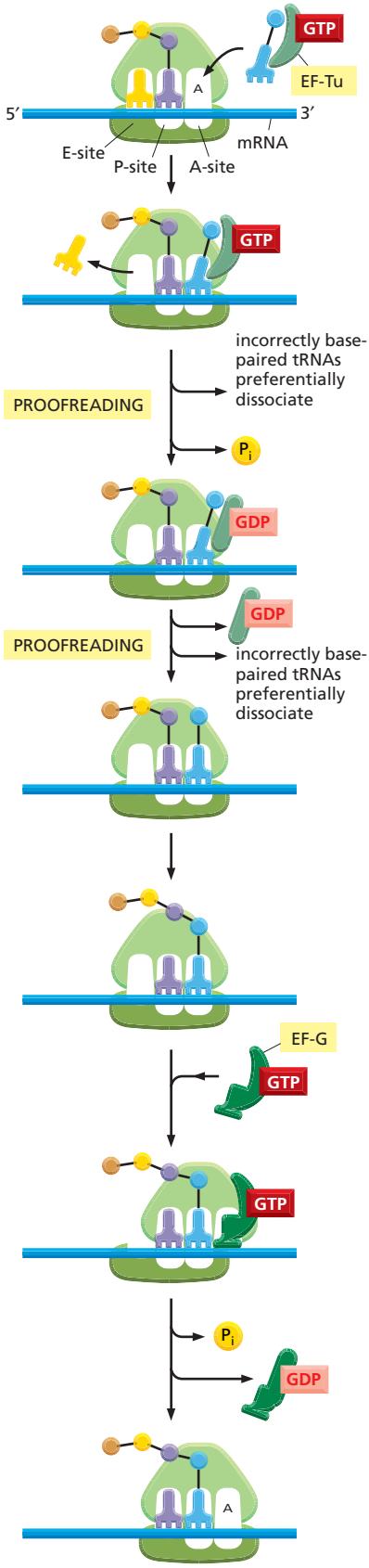


Figure 6–67 Detailed view of the translation cycle. The outline of translation presented in Figure 6–66 has been expanded to show the roles of two elongation factors EF-Tu and EF-G, which drive translation in the forward direction. As explained in the text, EF-Tu also provides two opportunities for proofreading of the codon–anticodon match. In this way, incorrectly paired tRNAs are selectively rejected, and the accuracy of translation is improved.

rapidly than those correctly bound because their interaction with the codon is weaker. Thus, most incorrectly bound tRNA molecules (as well as a significant number of correctly bound molecules) will leave the ribosome without being used for protein synthesis. All of these proofreading steps, taken together, are largely responsible for the 99.99% accuracy of the ribosome in translating RNA into protein.

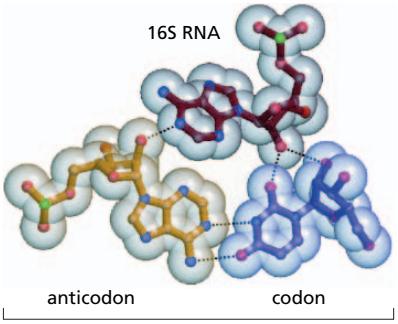


Figure 6–68 Recognition of correct codon–anticodon matches by the small subunit rRNA of the ribosome. Shown is the interaction between a nucleotide of the small subunit rRNA and the first nucleotide pair of a correctly paired codon–anticodon; similar interactions occur between other nucleotides of the rRNA and the second and third positions of the codon–anticodon pair. The small-subunit rRNA can form this network of hydrogen bonds only with correctly matched codon–anticodon pairs. As explained in the text, this codon–anticodon monitoring by the small-subunit rRNA increases the accuracy of protein synthesis. (From J.M. Ogle et al., *Science* 292:897–902, 2001. With permission from AAAS.)

The Ribosome Is a Ribozyme

The ribosome is a large complex composed of two-thirds RNA and one-third protein. The determination, in 2000, of the entire three-dimensional conformation of its large and small subunits is a major triumph of modern structural biology. The findings confirm earlier evidence that rRNAs—and not proteins—are responsible for the ribosome's overall structure, its ability to position tRNAs on the mRNA, and its catalytic activity in forming covalent peptide bonds. The ribosomal RNAs are folded into highly compact, precise three-dimensional structures that form the compact core of the ribosome and determine its overall shape (**Figure 6–69**).

In marked contrast to the central positions of the rRNAs, the ribosomal proteins are generally located on the surface and fill in the gaps and crevices of the folded RNA (**Figure 6–70**). Some of these proteins send out extended regions of polypeptide chain that penetrate short distances into holes in the RNA core (**Figure 6–71**). The main role of the ribosomal proteins seems to be to stabilize the RNA core, while permitting the changes in rRNA conformation that are necessary for this RNA to catalyze efficient protein synthesis. The proteins probably also aid in the initial assembly of the rRNAs that make up the core of the ribosome.

Not only are the A-, P-, and E-binding sites for tRNAs formed primarily by ribosomal RNAs, but the catalytic site for peptide bond formation is also formed by RNA, as the nearest amino acid is located more than 1.8 nm away.

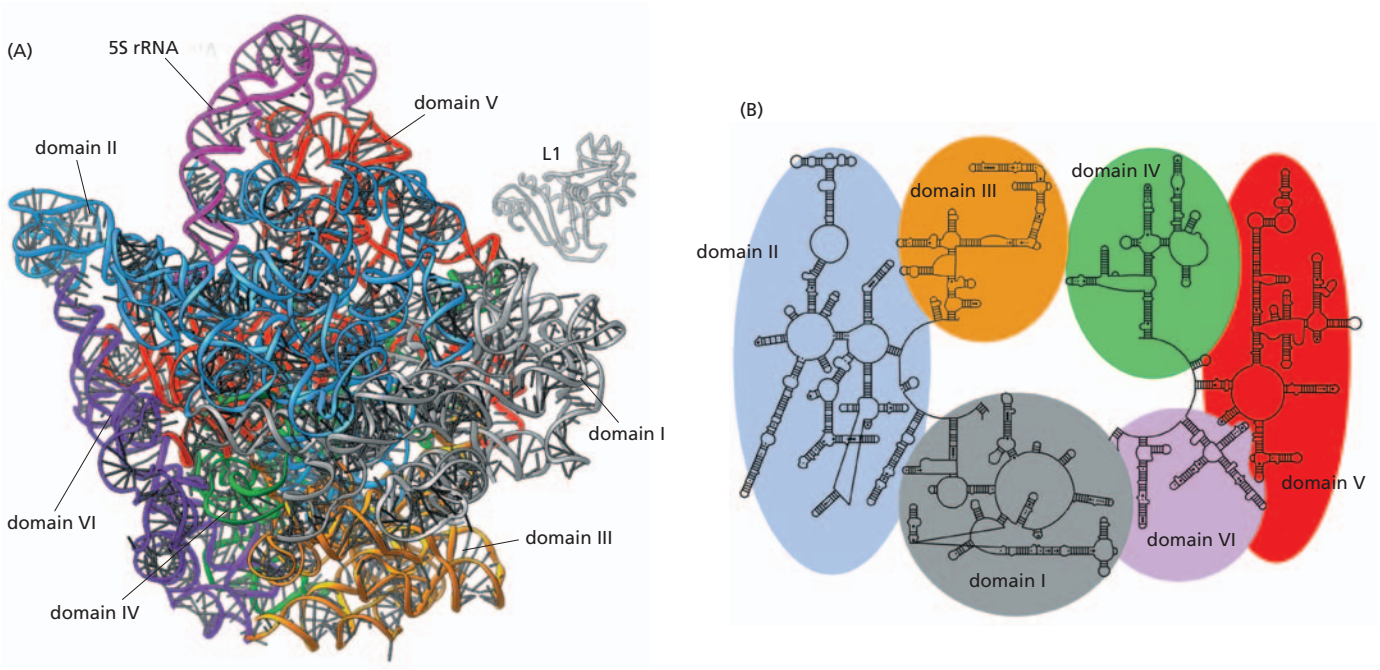


Figure 6–69 Structure of the rRNAs in the large subunit of a bacterial ribosome, as determined by x-ray crystallography. (A) Three-dimensional conformations of the large-subunit rRNAs (5S and 23S) as they appear in the ribosome. One of the protein subunits of the ribosome (L1) is also shown as a reference point, since it forms a characteristic protrusion on the ribosome. (B) Schematic diagram of the secondary structure of the 23S rRNA, showing the extensive network of base-pairing. The structure has been divided into six "domains" whose colors correspond to those in (A). The secondary-structure diagram is highly schematized to represent as much of the structure as possible in two dimensions. To do this, several discontinuities in the RNA chain have been introduced, although in reality the 23S RNA is a single RNA molecule. For example, the base of Domain III is continuous with the base of Domain IV even though a gap appears in the diagram. (Adapted from N. Ban et al., *Science* 289:905–920, 2000. With permission from AAAS.)

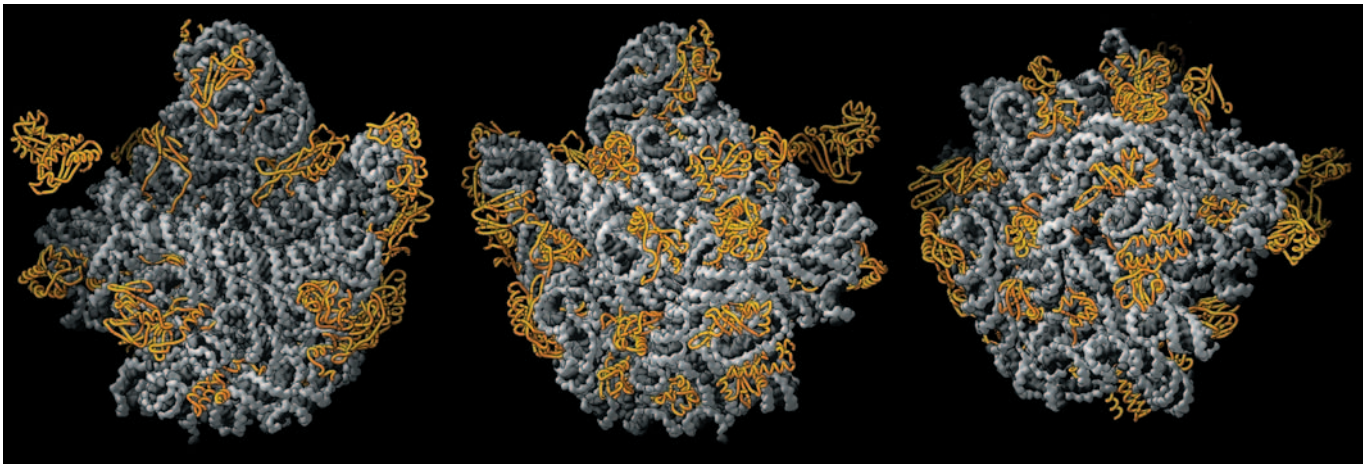


Figure 6–70 Location of the protein components of the bacterial large ribosomal subunit. The rRNAs (5S and 23S) are shown in *gray* and the large-subunit proteins (27 of the 31 total) in *gold*. For convenience, the protein structures depict only the polypeptide backbones. (A) Interface with the small subunit, the same view shown in Figure 6–64B. (B) Side opposite to that shown in (A), obtained by rotating (A) by 180° around a vertical axis. (C) Further slight rotation of (B) through a diagonal axis, allowing a view into the peptide exit channel in the center of the structure. (From N. Ban et al., *Science* 289:905–920, 2000. With permission from AAAS.)

This discovery came as a surprise to biologists because, unlike proteins, RNA does not contain easily ionizable functional groups that can be used to catalyze sophisticated reactions like peptide bond formation. Moreover, metal ions, which are often used by RNA molecules to catalyze chemical reactions (as discussed later in the chapter), were not observed at the active site of the ribosome. Instead, it is believed that the 23S rRNA forms a highly structured pocket that, through a network of hydrogen bonds, precisely orients the two reactants (the growing peptide chain and an aminoacyl-tRNA) and thereby greatly accelerates their covalent joining. In addition, the tRNA in the P site contributes to the active site, perhaps supplying a functional OH group that participates directly in the catalysis. This mechanism may ensure that catalysis occurs only when the tRNA is properly positioned in the ribosome.

RNA molecules that possess catalytic activity are known as **ribozymes**. We saw earlier in this chapter how other ribozymes function in self-splicing reactions (for example, see Figure 6–36). In the final section of this chapter, we consider what the ability of RNA molecules to function as catalysts for a wide variety of different reactions might mean for the early evolution of living cells. For now, we merely note that there is good reason to suspect that RNA rather than protein molecules served as the first catalysts for living cells. If so, the ribosome, with its RNA core, may be a relic of an earlier time in life's history—when protein synthesis evolved in cells that were run almost entirely by ribozymes.

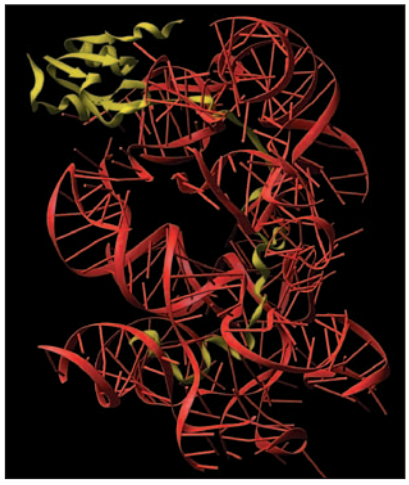


Figure 6–71 Structure of the L15 protein in the large subunit of the bacterial ribosome. The globular domain of the protein lies on the surface of the ribosome and an extended region penetrates deeply into the RNA core of the ribosome. The L15 protein is shown in *yellow* and a portion of the ribosomal RNA core is shown in *red*. (From D. Klein, P.B. Moore and T.A. Steitz, *J. Mol. Biol.* 340:141–147, 2004. With permission from Academic Press.)

Nucleotide Sequences in mRNA Signal Where to Start Protein Synthesis

The initiation and termination of translation share features of the translation elongation cycle described above. The site at which protein synthesis begins on the mRNA is especially crucial, since it sets the reading frame for the whole length of the message. An error of one nucleotide either way at this stage would cause every subsequent codon in the message to be misread, resulting in a nonfunctional protein with a garbled sequence of amino acids. The initiation step is also important because for most genes it is the last point at which the cell can decide whether the mRNA is to be translated and the protein synthesized; the rate of initiation is thus one determinant of the rate at which any protein is synthesized. We shall see in Chapter 7 that cells use several mechanisms to regulate translation initiation.

The translation of an mRNA begins with the codon AUG, and a special tRNA is required to start translation. This **initiator tRNA** always carries the amino acid methionine (in bacteria, a modified form of methionine—formylmethionine—is used), with the result that all newly made proteins have methionine as the first amino acid at their N-terminus, the end of a protein that is synthesized first. This methionine is usually removed later by a specific protease. The initiator tRNA can be specially recognized by initiation factors because it has a nucleotide sequence distinct from that of the tRNA that normally carries methionine.

In eucaryotes, the initiator tRNA–methionine complex (Met–tRNAi) is first loaded into the small ribosomal subunit along with additional proteins called **eucaryotic initiation factors**, or **eIFs** (**Figure 6–72**). Of all the aminoacyl-tRNAs in the cell, only the methionine-charged initiator tRNA is capable of tightly binding the small ribosome subunit without the complete ribosome being present and it binds directly to the P-site. Next, the small ribosomal subunit binds to the 5′ end of an mRNA molecule, which is recognized by virtue of its 5′ cap and its two bound initiation factors, eIF4E (which directly binds the cap) and eIF4G (see Figure 6–40). The small ribosomal subunit then moves forward (5′ to 3′) along the mRNA, searching for the first AUG. Additional initiation factors that act as ATP-powered helicases facilitate the ribosome's movement through RNA secondary structure. In 90% of mRNAs, translation begins at the first AUG encountered by the small subunit. At this point, the initiation factors dissociate, allowing the large ribosomal subunit to assemble with the complex and complete the ribosome. The initiator tRNA is still bound to the P-site, leaving the A-site vacant. Protein synthesis is therefore ready to begin (see Figure 6–72).

The nucleotides immediately surrounding the start site in eucaryotic mRNAs influence the efficiency of AUG recognition during the above scanning process. If this recognition site differs substantially from the consensus recognition sequence (5′-ACCAUGG-3′), scanning ribosomal subunits will sometimes ignore the first AUG codon in the mRNA and skip to the second or third AUG codon instead. Cells frequently use this phenomenon, known as "leaky scanning," to produce two or more proteins, differing in their N-termini, from the same mRNA molecule. It allows some genes to produce the same protein with and without a signal sequence attached at its N-terminus, for example, so that the protein is directed to two different compartments in the cell.

The mechanism for selecting a start codon in bacteria is different. Bacterial mRNAs have no 5′ caps to signal the ribosome where to begin searching for the start of translation. Instead, each bacterial mRNA contains a specific ribosome-binding site (called the Shine–Dalgarno sequence, named after its discoverers) that is located a few nucleotides upstream of the AUG at which translation is to begin. This nucleotide sequence, with the consensus 5′-AGGAGGU-3′, forms base pairs with the 16S rRNA of the small ribosomal subunit to position the initiating AUG codon in the ribosome. A set of translation initiation factors orchestrates this interaction, as well as the subsequent assembly of the large ribosomal subunit to complete the ribosome.

Unlike a eucaryotic ribosome, a bacterial ribosome can therefore readily assemble directly on a start codon that lies in the interior of an mRNA molecule, so long as a ribosome-binding site precedes it by several nucleotides. As a result, bacterial mRNAs are often *polycistronic*—that is, they encode several different proteins, each of which is translated from the same mRNA molecule (**Figure 6–73**). In contrast, a eucaryotic mRNA generally encodes only a single protein.

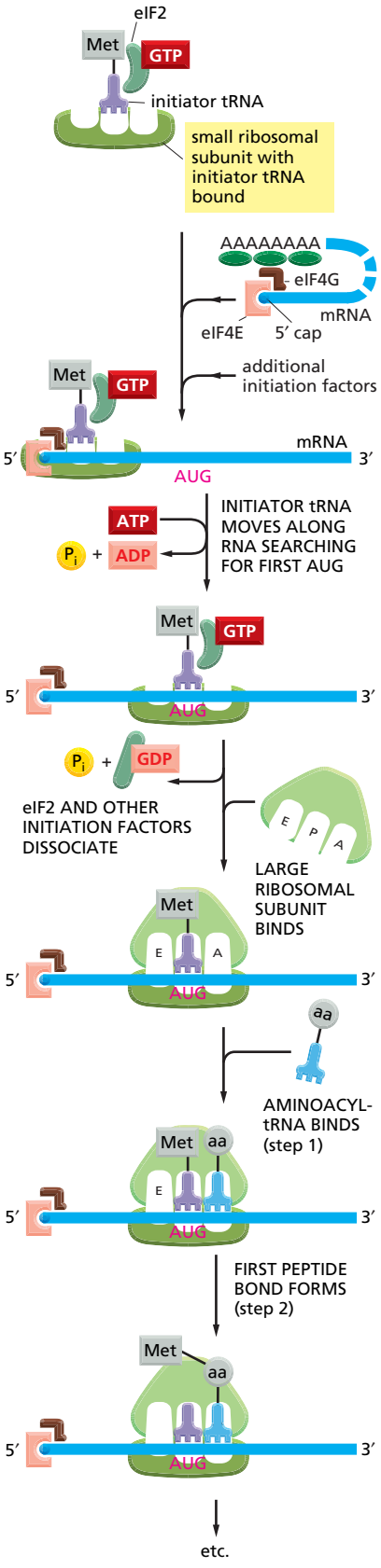


Figure 6–72 The initiation of protein synthesis in eucaryotes. Only three of the many translation initiation factors required for this process are shown. Efficient translation initiation also requires the poly-A tail of the mRNA bound by poly-A-binding proteins which, in turn, interact with eIF4G. In this way, the translation apparatus ascertains that both ends of the mRNA are intact before initiating protein synthesis (see Figure 6–40). Although only one GTP hydrolysis event is shown in the figure, a second is known to occur just before the large and small ribosomal subunits join.

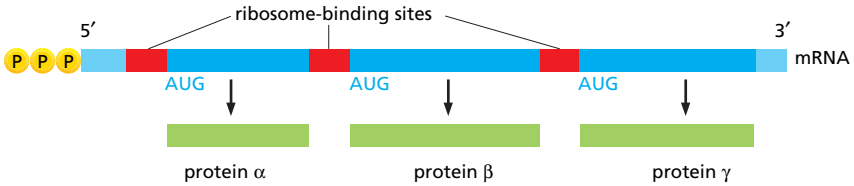


Figure 6–73 Structure of a typical bacterial mRNA molecule. Unlike eucaryotic ribosomes, which typically require a capped 5′ end, procaryotic ribosomes initiate transcription at ribosome-binding sites (Shine–Dalgarno sequences), which can be located anywhere along an mRNA molecule. This property of ribosomes permits bacteria to synthesize more than one type of protein from a single mRNA molecule.

Stop Codons Mark the End of Translation

The end of the protein-coding message is signaled by the presence of one of three *stop codons* (UAA, UAG, or UGA) (see Figure 6–50). These are not recognized by a tRNA and do not specify an amino acid, but instead signal to the ribosome to stop translation. Proteins known as *release factors* bind to any ribosome with a stop codon positioned in the A site, forcing the peptidyl transferase in the ribosome to catalyze the addition of a water molecule instead of an amino acid to the peptidyl-tRNA (**Figure 6–74**). This reaction frees the carboxyl end of the growing polypeptide chain from its attachment to a tRNA molecule, and since only this attachment normally holds the growing polypeptide to the ribosome, the completed protein chain is immediately released into the cytoplasm. The ribosome then releases the mRNA and separates into the large and small subunits, which can assemble on this or another mRNA molecule to begin a new round of protein synthesis.

Release factors are an example of *molecular mimicry*, whereby one type of macromolecule resembles the shape of a chemically unrelated molecule. In this case, the three-dimensional structure of release factors (made entirely of protein) resembles the shape and charge distribution of a tRNA molecule (**Figure 6–75**). This shape and charge mimicry helps them enter the A-site on the ribosome and cause translation termination.

During translation, the nascent polypeptide moves through a large, water-filled tunnel (approximately 10 nm × 1.5 nm) in the large subunit of the ribosome (see Figure 6–70C). The walls of this tunnel, made primarily of 23S rRNA, are a patchwork of tiny hydrophobic surfaces embedded in a more extensive hydrophilic surface. This structure is not complementary to any peptide, and thus provides a "Teflon" coating through which a polypeptide chain can easily slide. The dimensions of the tunnel suggest that nascent proteins are largely unstructured as they pass through the ribosome, although some α-helical regions of the protein can form before leaving the ribosome tunnel. As it leaves the ribosome, a newly synthesized protein must fold into its proper three-dimensional conformation to be useful to the cell, and later in this chapter we discuss how this folding occurs. First, however, we describe several additional aspects of the translation process itself.

Proteins Are Made on Polyribosomes

The synthesis of most protein molecules takes between 20 seconds and several minutes. During this very short period, however, it is usual for multiple initiations to take place on each mRNA molecule being translated. As soon as the preceding ribosome has translated enough of the nucleotide sequence to move out

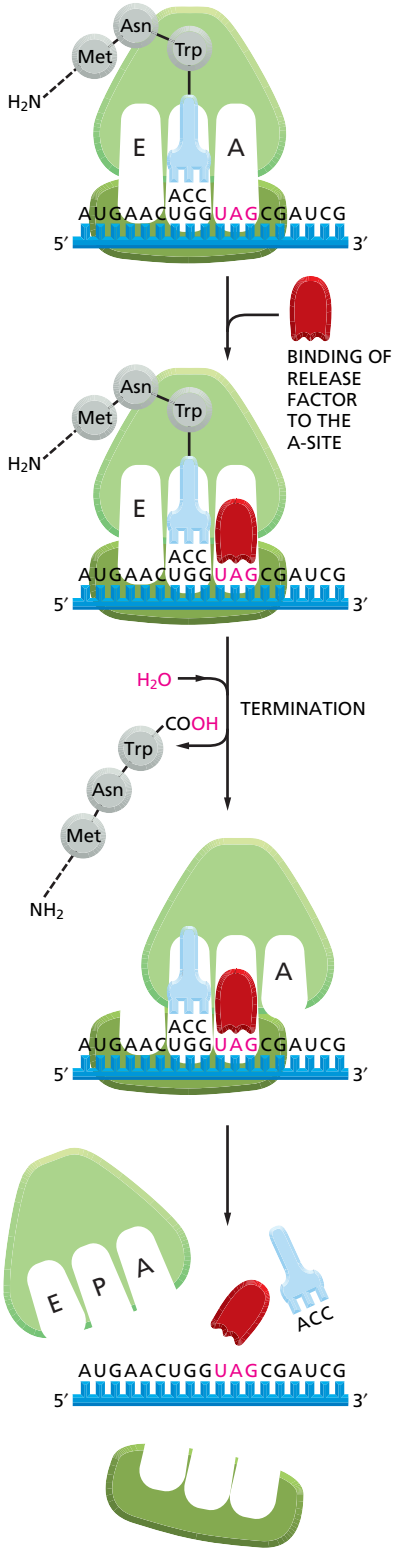


Figure 6–74 The final phase of protein synthesis. The binding of a release factor to an A-site bearing a stop codon terminates translation. The completed polypeptide is released and, in a series of reactions that requires additional proteins and GTP hydrolysis (not shown), the ribosome dissociates into its two separate subunits.

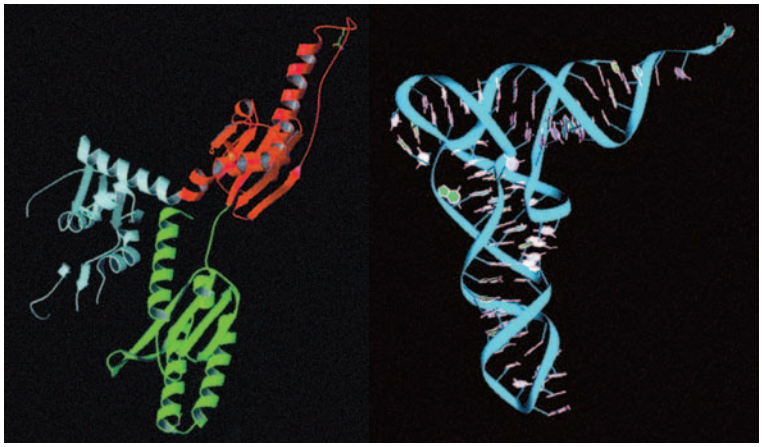


Figure 6–75 The structure of a human translation release factor (eRF1) and its resemblance to a tRNA molecule. The protein is on the *left* and the tRNA on the *right*. (From H. Song et al., *Cell* 100:311–321, 2000. With permission from Elsevier.)

of the way, the 5′ end of the mRNA is threaded into a new ribosome. The mRNA molecules being translated are therefore usually found in the form of *polyribosomes* (or *polysomes*): large cytoplasmic assemblies made up of several ribosomes spaced as close as 80 nucleotides apart along a single mRNA molecule (**Figure 6–76**). These multiple initiations allow the cell to make many more protein molecules in a given time than would be possible if each had to be completed before the next could start. <GAAG>

Both bacteria and eucaryotes use polysomes, and both employ additional strategies to speed up the overall rate of protein synthesis even further. Because bacterial mRNA does not need to be processed and is accessible to ribosomes while it is being made, ribosomes can attach to the free end of a bacterial mRNA molecule and start translating it even before the transcription of that RNA is complete, following closely behind the RNA polymerase as it moves along DNA. In eucaryotes, as we have seen, the 5′ and 3′ ends of the mRNA interact (see Figures 6–40 and 6–76A); therefore, as soon as a ribosome dissociates, its two subunits are in an optimal position to reinitiate translation on the same mRNA molecule.

There Are Minor Variations in the Standard Genetic Code

As discussed in Chapter 1, the genetic code (shown in Figure 6–50) applies to all three major branches of life, providing important evidence for the common

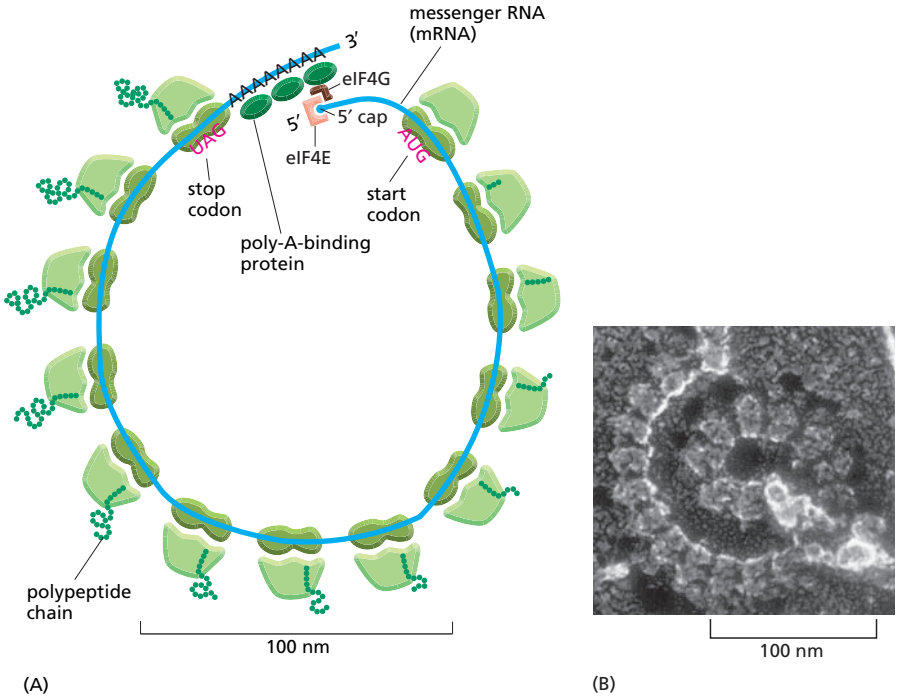


Figure 6–76 A polyribosome. (A) Schematic drawing showing how a series of ribosomes can simultaneously translate the same eucaryotic mRNA molecule. (B) Electron micrograph of a polyribosome from a eucaryotic cell. (B, courtesy of John Heuser.)

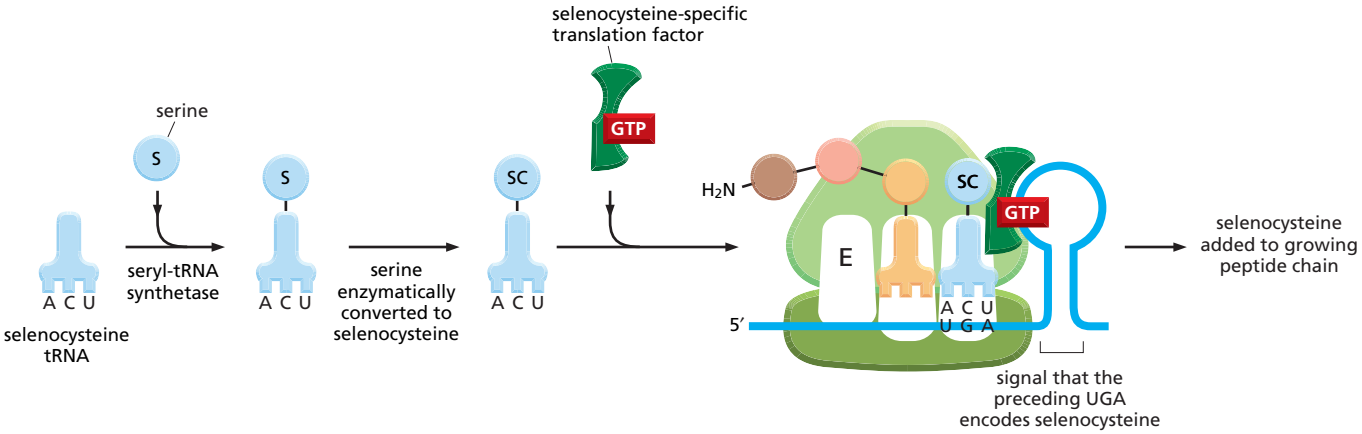


Figure 6–77 Incorporation of selenocysteine into a growing polypeptide chain. A specialized tRNA is charged with serine by the normal seryl-tRNA synthetase, and the serine is subsequently converted enzymatically to selenocysteine. A specific RNA structure in the mRNA (a stem and loop structure with a particular nucleotide sequence) signals that selenocysteine is to be inserted at the neighboring UGA codon. As indicated, this event requires the participation of a selenocysteine-specific translation factor.

ancestry of all life on Earth. Although rare, there are exceptions to this code. For example, *Candida albicans*, the most prevalent human fungal pathogen, translates the codon CUG as serine, whereas nearly all other organisms translate it as leucine. Mitochondria (which have their own genomes and encode much of their translational apparatus) often deviate from the standard code. For example, in mammalian mitochondria AUA is translated as methionine, whereas in the cytosol of the cell it is translated as isoleucine (see Table 14–3, p. 862). This type of deviation in the genetic code is "hardwired" into the organisms or the organelles in which it occurs.

A different type of variation, sometimes called *translation recoding*, occurs in many cells. In this case, other nucleotide sequence information present in an mRNA can change the meaning of the genetic code at a particular site in the mRNA molecule. The standard code allows cells to manufacture proteins using only 20 amino acids. However, bacteria, archaea, and eucaryotes have available to them a twenty-first amino acid that can be incorporated directly into a growing polypeptide chain through translation recoding. Selenocysteine, which is essential for the efficient function of a variety of enzymes, contains a selenium atom in place of the sulfur atom of cysteine. Selenocysteine is enzymatically produced from a serine attached to a special tRNA molecule that base-pairs with the UGA codon, a codon normally used to signal a translation stop. The mRNAs for proteins in which selenocysteine is to be inserted at a UGA codon carry an additional nucleotide sequence in the mRNA nearby that causes this recoding event (**Figure 6–77**).

Another form of recoding, *translational frameshifting*, allows more than one protein to be synthesized from a single mRNA. Retroviruses, members of a large group of eucaryotic-infecting pathogens, commonly use translational frameshifting to make both the capsid proteins *(Gag proteins)* and the viral reverse transcriptase and integrase *(Pol proteins)* from the same RNA transcript (see Figure 5–73). The virus needs many more copies of the Gag proteins than it does of the Pol proteins. This quantitative adjustment is achieved by encoding the *Pol* genes just after the *Gag* genes but in a different reading frame. Small amounts of the Pol gene products are made because, on occasion, an upstream translational frameshift allows the Gag protein stop codon to be bypassed. This frameshift occurs at a particular codon in the mRNA and requires a specific *recoding signal*, which seems to be a structural feature of the RNA sequence downstream of this site (**Figure 6–78**).

Inhibitors of Procaryotic Protein Synthesis Are Useful as Antibiotics

Many of the most effective antibiotics used in modern medicine are compounds made by fungi that inhibit bacterial protein synthesis. Fungi and bacteria compete for many of the same environmental niches, and millions of years of coevolution has resulted in fungi producing potent bacterial inhibitors. Some of these

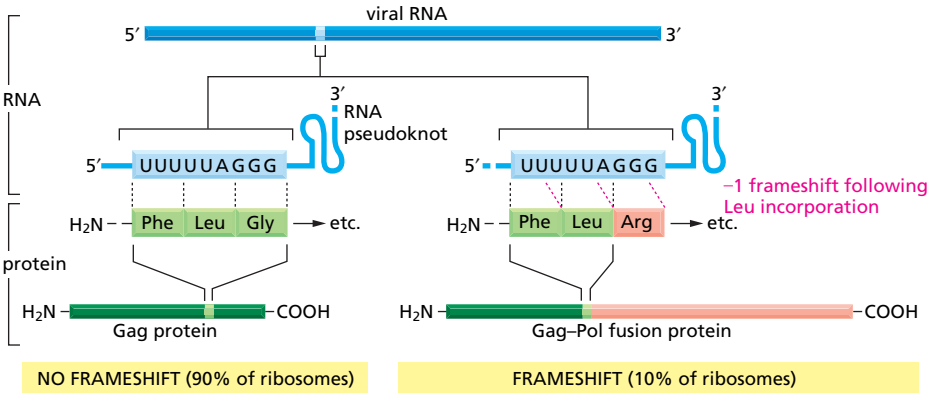


Figure 6–78 The translational frameshifting that produces the reverse transcriptase and integrase of a retrovirus. The viral reverse transcriptase and integrase are produced by proteolytic processing of a large protein (the Gag–Pol fusion protein) consisting of both the Gag and Pol amino acid sequences. Proteolytic processing of the more abundant Gag protein produces the viral capsid proteins. Both the Gag and the Gag–Pol fusion proteins start with identical mRNA, but whereas the Gag protein terminates at a stop codon downstream of the sequence shown, translation of the Gag–Pol fusion protein bypasses this stop codon, allowing the synthesis of the longer Gag–Pol fusion protein. The stop-codon-bypass is made possible by a controlled translational frameshift, as illustrated. Features in the local RNA structure (including the RNA loop shown) cause the $tRNA^{Leu}$ attached to the C-terminus of the growing polypeptide chain occasionally to slip backward by one nucleotide on the ribosome, so that it pairs with a UUU codon instead of the UUA codon that had initially specified its incorporation; the next codon (AGG) in the new reading frame specifies an arginine rather than a glycine. This controlled slippage is due in part to a *pseudoknot* that forms in the viral mRNA (see Figure 6–102). The sequence shown is from the human AIDS virus, HIV. (Adapted from T. Jacks et al., *Nature* 331:280–283, 1988. With permission from Macmillan Publishers Ltd.)

drugs exploit the structural and functional differences between bacterial and eucaryotic ribosomes so as to interfere preferentially with the function of bacterial ribosomes. Thus humans can take high dosages of some of these compounds without undue toxicity. Many antibiotics lodge in pockets in the ribosomal RNAs and simply interfere with the smooth operation of the ribosome (**Figure 6–79**). **Table 6–4** lists some of the more common antibiotics of this kind along with several other inhibitors of protein synthesis, some of which act on eucaryotic cells and therefore cannot be used as antibiotics.

Because they block specific steps in the processes that lead from DNA to protein, many of the compounds listed in Table 6–4 are useful for cell biological studies. Among the most commonly used drugs in such investigations are *chloramphenicol*, *cycloheximide*, and *puromycin*, all of which specifically inhibit protein synthesis. In a eucaryotic cell, for example, chloramphenicol inhibits protein synthesis on ribosomes only in mitochondria (and in chloroplasts in plants), presumably reflecting the procaryotic origins of these organelles (discussed in Chapter 14). Cycloheximide, in contrast, affects only ribosomes in the cytosol. Puromycin is especially interesting because it is a structural analog of a tRNA molecule linked to an amino acid and is therefore another example of molecular mimicry; the ribosome mistakes it for an authentic amino acid and covalently incorporates it at the C-terminus of the growing peptide chain, thereby causing the premature termination and release of the polypeptide. As might be expected, puromycin inhibits protein synthesis in both procaryotes and eucaryotes.

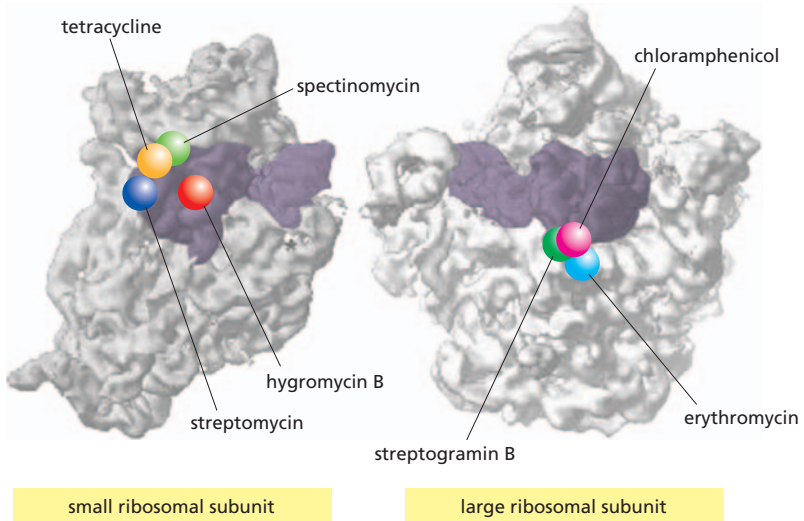


Figure 6–79 Binding sites for antibiotics on the bacterial ribosome. The small *(left)* and large *(right)* subunits of the ribosome are arranged as though the ribosome has been opened like a book; the bound tRNA molecules are shown in *purple* (see Figure 6–64). Most of the antibiotics shown bind directly to pockets formed by the ribosomal RNA molecules. Hygromycin B induces errors in translation, spectinomycin blocks the translocation of the peptidyl-tRNA from the A-site to the P-site, and streptogramin B prevents elongation of nascent peptides. Table 6–4 lists the inhibitory mechanisms of the other antibiotics shown in the figure. (Adapted from J. Poehlsgaard and S. Douthwaite, *Nat. Rev. Microbiol.* 3:870–881, 2005. With permission from Macmillan Publishers Ltd.)

Table 6–4 Inhibitors of Protein or RNA Synthesis

INHIBITOR	SPECIFIC EFFECT
Acting only on bacteria	
Tetracycline	blocks binding of aminoacyl-tRNA to A-site of ribosome
Streptomycin	prevents the transition from translation initiation to chain elongation and also causes miscoding
Chloramphenicol	blocks the peptidyl transferase reaction on ribosomes (step 2 in Figure 6–66)
Erythromycin	binds in the exit channel of the ribosome and thereby inhibits elongation of the peptide chain
Rifamycin	blocks initiation of RNA chains by binding to RNA polymerase (prevents RNA synthesis)
Acting on bacteria and eucaryotes	
Puromycin	causes the premature release of nascent polypeptide chains by its addition to the growing chain end
Actinomycin D	binds to DNA and blocks the movement of RNA polymerase (prevents RNA synthesis)
Acting on eucaryotes but not bacteria	
Cycloheximide	blocks the translocation reaction on ribosomes (step 3 in Figure 6–66)
Anisomycin	blocks the peptidyl transferase reaction on ribosomes (step 2 in Figure 6–66)
α-Amanitin	blocks mRNA synthesis by binding preferentially to RNA polymerase II

The ribosomes of eucaryotic mitochondria (and chloroplasts) often resemble those of bacteria in their sensitivity to inhibitors. Therefore, some of these antibiotics can have a deleterious effect on human mitochondria.

Accuracy in Translation Requires the Expenditure of Free Energy

Translation by the ribosome is a compromise between the opposing constraints of accuracy and speed. We have seen, for example, that the accuracy of translation (1 mistake per 10^4 amino acids joined) requires time delays each time a new amino acid is added to a growing polypeptide chain, producing an overall speed of translation of 20 amino acids incorporated per second in bacteria. Mutant bacteria with a specific alteration in the small ribosomal subunit have longer delays and translate mRNA into protein with an accuracy considerably higher than this; however, protein synthesis is so slow in these mutants that the bacteria are barely able to survive.

We have also seen that attaining the observed accuracy of protein synthesis requires the expenditure of a great deal of free energy; this is expected, since, as discussed in Chapter 2, there is a price to be paid for any increase in order in the cell. In most cells, protein synthesis consumes more energy than any other biosynthetic process. At least four high-energy phosphate bonds are split to make each new peptide bond: two are consumed in charging a tRNA molecule with an amino acid (see Figure 6–56), and two more drive steps in the cycle of reactions occurring on the ribosome during synthesis itself (see Figure 6–67). In addition, extra energy is consumed each time that an incorrect amino acid linkage is hydrolyzed by a tRNA synthetase (see Figure 6–59) and each time that an incorrect tRNA enters the ribosome, triggers GTP hydrolysis, and is rejected (see Figure 6–67). To be effective, these proofreading mechanisms must also allow an appreciable fraction of correct interactions to be removed; for this reason, proofreading is even more costly in energy than it might seem.

Quality Control Mechanisms Act to Prevent Translation of Damaged mRNAs

In eucaryotes, mRNA production involves both transcription and a series of elaborate RNA-processing steps; these take place in the nucleus, segregated from ribosomes, and only when the processing is complete are the mRNAs transported to the cytoplasm to be translated (see Figure 6–40). However, this scheme is not foolproof, and some incorrectly processed mRNAs are inadvertently sent to the cytoplasm. In addition, mRNAs that were flawless when they left the nucleus can become broken or otherwise damaged in the cytosol. The danger of translating damaged or incompletely processed mRNAs (which would produce truncated or otherwise aberrant proteins) is apparently so great that the cell has several backup measures to prevent this from happening.

To avoid translating broken mRNAs, the 5′ cap and the poly-A tail are both recognized by the translation-initiation machinery before translation begins (see Figure 6–72). To help ensure that mRNAs are properly spliced before they are translated, the exon junction complex (EJC), which is deposited on the mRNA following splicing (see Figure 6–40), stimulates the subsequent translation of the mRNA.

But the most powerful mRNA surveillance system, called **nonsense-mediated mRNA decay**, eliminates defective mRNAs before they can be efficiently translated into protein. This mechanism is brought into play when the cell determines that an mRNA molecule has a nonsense (stop) codon (UAA, UAG, or UGA) in the "wrong" place—a situation likely to arise in an mRNA molecule that has been improperly spliced. Aberrant splicing will usually result in the random introduction of a nonsense codon into the reading frame of the mRNA, especially in organisms, such as humans, that have a large average intron size (see Figure 6–32B).

This surveillance mechanism begins as an mRNA molecule is being transported from the nucleus to the cytosol. As its 5′ end emerges from the nuclear pore, the mRNA is met by a ribosome, which begins to translate it. As translation proceeds, the exon junction complexes (EJC) bound to the mRNA at each splice-site are apparently displaced by the moving ribosome. The normal stop codon will be within the last exon, so by the time the ribosome reaches it and stalls, no more EJCs should be bound to the mRNA. If this is the case, the mRNA "passes inspection" and is released to the cytosol where it can be translated in earnest (**Figure 6–80**). However, if the ribosome reaches a premature stop codon and stalls, it senses that EJCs remain and the bound mRNA molecule is rapidly degraded. In this way, the first round of translation allows the cell to test the fitness of each mRNA molecule as it exits the nucleus.

Nonsense-mediated decay may have been especially important in evolution, allowing eucaryotic cells to more easily explore new genes formed by DNA rearrangements, mutations, or alternative patterns of splicing—by selecting only those mRNAs for translation that can produce a full-length protein. Nonsense-mediated decay is also important in cells of the developing immune system, where the extensive DNA rearrangements that occur (see Figure 25–36) often generate premature termination codons. The surveillance system degrades the mRNAs produced from such rearranged genes, thereby avoiding the potential toxic effects of truncated proteins.

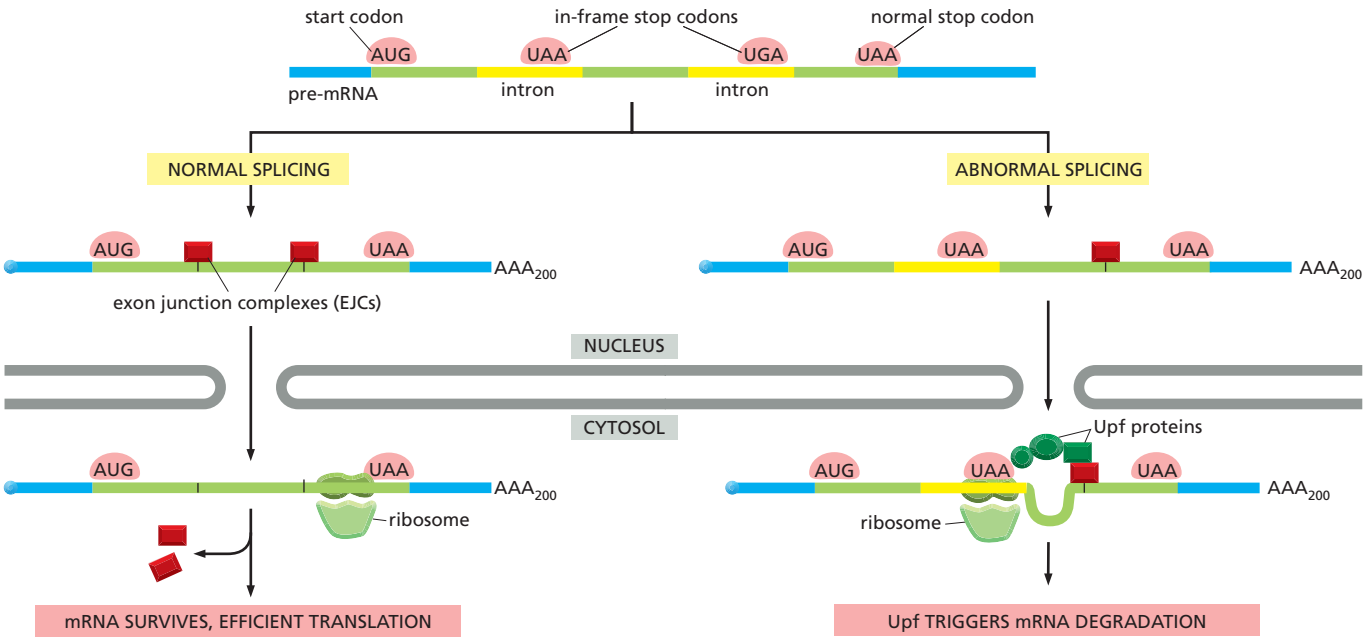


Figure 6–80 Nonsense-mediated mRNA decay. As shown on the right, the failure to correctly splice a pre-mRNA often introduces a premature stop codon into the reading frame for the protein. The introduction of such an "in-frame" stop codon is particularly likely to occur in mammals, where the introns tend to be very long. When translated, these abnormal mRNAs produce aberrant proteins, which could damage the cell. However, as shown at the bottom right of the figure, these abnormal RNAs are destroyed by the nonsense-mediated decay mechanism. According to one model, an mRNA molecule, bearing exon junction complexes (EJCs) to mark successfully completed splices, is first met by a ribosome that performs a "test" round of translation. As the mRNA passes through the tight channel of the ribosome, the EJCs are stripped off, and successful mRNAs are released to undergo multiple rounds of translation *(left side)*. However, if an in-frame stop codon is encountered before the final exon junction complex is reached (*right side*), the mRNA undergoes nonsense-mediated decay, which is triggered by the Upf proteins (*green*) that bind to each EJC. Note that, to trigger nonsense-mediated decay, the premature stop codon must be in the same reading frame as that of the normal protein. (Adapted from J. Lykke-Andersen et al., *Cell* 103:1121–1131, 2000. With permission from Elsevier.)

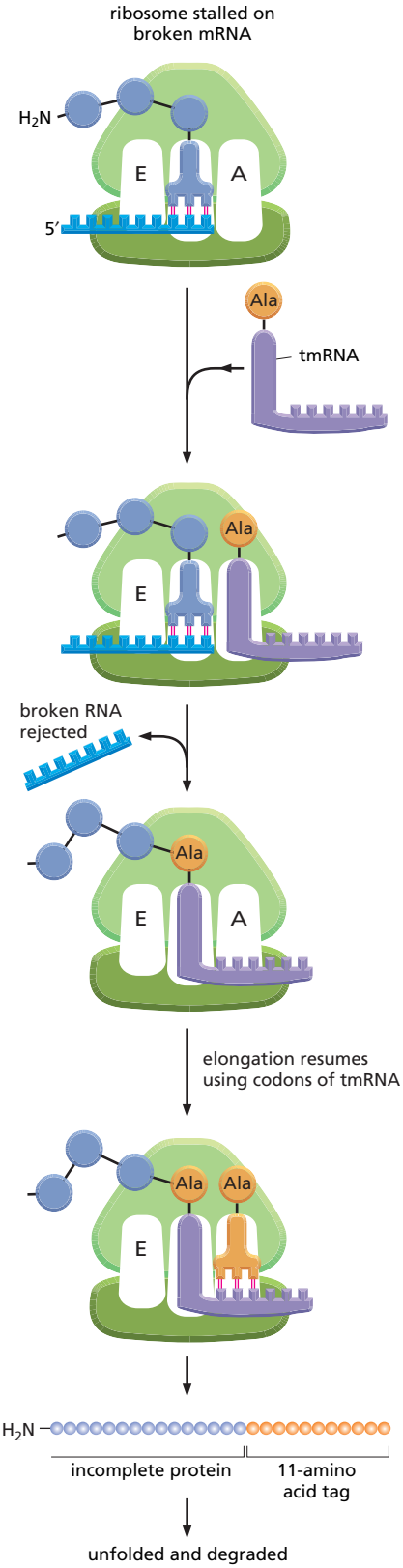


Figure 6–81 The rescue of a bacterial ribosome stalled on an incomplete mRNA molecule. The tmRNA shown is a 363-nucleotide RNA with both tRNA and mRNA functions, hence its name. It carries an alanine and can enter the vacant A-site of a stalled ribosome to add this alanine to a polypeptide chain, mimicking a tRNA although no codon is present to guide it. The ribosome then translates 10 codons from the tmRNA, completing an 11 amino acid tag on the protein. Proteases recognize this tag and degrade the entire protein. Although the example shown in the figure is from bacteria, eucaryotes can employ a similar strategy.

Finally, the nonsense-mediated surveillance pathway plays an important role in mitigating the symptoms of many inherited human diseases. As we have seen, inherited diseases are usually caused by mutations that spoil the function of a key protein, such as hemoglobin or one of the blood clotting factors. Approximately one-third of all genetic disorders in humans result from nonsense mutations or mutations (such as frameshift mutations or splice-site mutations) that place nonsense mutations into the gene's reading frame. In individuals that carry one mutant and one functional gene, nonsense-mediated decay eliminates the aberrant mRNA and thereby prevents a potentially toxic protein from being made. Without this safeguard, individuals with one functional and one mutant "disease gene" would likely suffer much more severe symptoms.

We saw earlier in this chapter that bacteria lack the elaborate mRNA processing found in eucaryotes and that translation often begins before the synthesis of the RNA molecule is completed. Yet bacteria also have quality control mechanisms to deal with incompletely synthesized and broken mRNAs. When the bacterial ribosome translates to the end of an incomplete RNA it stalls and does not release the RNA. Rescue comes in the form of a special RNA (called tmRNA), which enters the A-site of the ribosome and is itself translated, releasing the ribosome. The special 11 amino acid tag thus added to the C-terminus of the truncated protein signals to proteases that the entire protein is to be degraded (**Figure 6–81**).

Some Proteins Begin to Fold While Still Being Synthesized

The process of gene expression is not over when the genetic code has been used to create the sequence of amino acids that constitutes a protein. To be useful to the cell, this new polypeptide chain must fold up into its unique three-dimensional conformation, bind any small-molecule cofactors required for its activity, be appropriately modified by protein kinases or other protein-modifying enzymes, and assemble correctly with the other protein subunits with which it functions (**Figure 6–82**).

The information needed for all of the steps listed above is ultimately contained in the sequence of linked amino acids that the ribosome produces when it translates an mRNA molecule into a polypeptide chain. As discussed in Chapter 3, when a protein folds into a compact structure, it buries most of its hydrophobic residues in an interior core. In addition, large numbers of noncovalent interactions form between various parts of the molecule. It is the sum of all of these energetically favorable arrangements that determines the final folding pattern of the polypeptide chain—as the conformation of lowest free energy (see p. 130).

Through many millions of years of evolution, the amino acid sequence of each protein has been selected not only for the conformation that it adopts but also for an ability to fold rapidly. For some proteins, this folding begins immediately, as the protein spins out of the ribosome, starting from the N-terminal end. In these cases, as each protein domain emerges from the ribosome, within a few seconds it forms a compact structure that contains most of the final secondary features (α helices and β sheets) aligned in roughly the right conformation (**Figure 6–83**). For many protein domains, this unusually dynamic and flexible state called a *molten globule*, is the starting point for a relatively slow process in which many side-chain adjustments occur that eventually form the correct tertiary

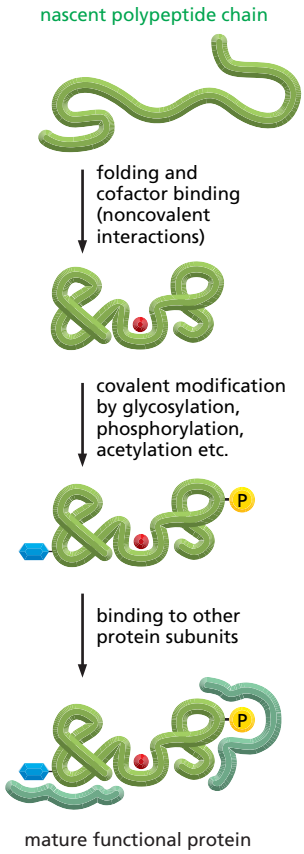


Figure 6–82 Steps in the creation of a functional protein. As indicated, translation of an mRNA sequence into an amino acid sequence on the ribosome is not the end of the process of forming a protein. To function, the completed polypeptide chain must fold correctly into its three-dimensional conformation, bind any cofactors required, and assemble with its partner protein chains (if any). Noncovalent bond formation drives these changes. As indicated, many proteins also require covalent modifications of selected amino acids. Although the most frequent modifications are protein glycosylation and protein phosphorylation, more than 100 different types of covalent modifications are known (see, for example, Figure 3–81).

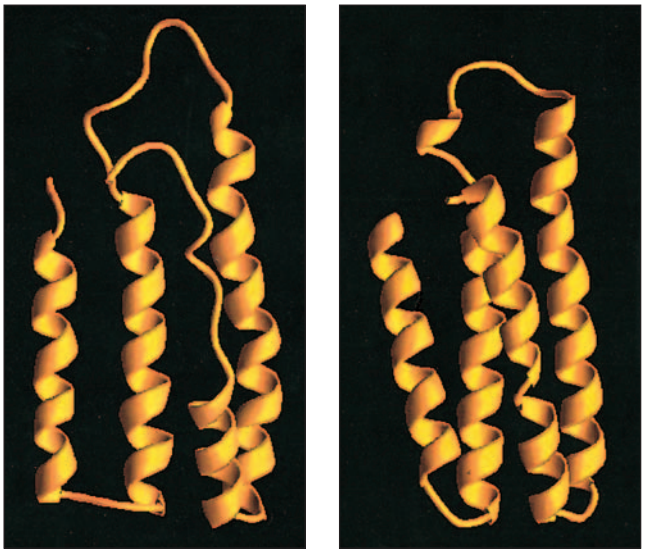
structure. It takes several minutes to synthesize a protein of average size, and for some proteins much of the folding process is complete by the time the ribosome releases the C-terminal end of a protein (**Figure 6–84**).

Molecular Chaperones Help Guide the Folding of Most Proteins

Most proteins probably do not begin to fold during their synthesis. Instead, they are met at the ribosome by a special class of proteins called **molecular chaperones**. Molecular chaperones are useful for cells because there are many different paths that can be taken to convert an unfolded or partially folded protein to its final compact conformation. For many proteins, some of the intermediates formed along the way would aggregate and be left as off-pathway dead ends without the intervention of a chaperone (**Figure 6–85**).

Many molecular chaperones are called *heat-shock proteins* (designated *Hsp*), because they are synthesized in dramatically increased amounts after a brief exposure of cells to an elevated temperature (for example, 42°C for cells that normally live at 37°C). This reflects the operation of a feedback system that responds to an increase in misfolded proteins (such as those produced by elevated temperatures) by boosting the synthesis of the chaperones that help these proteins refold.

There are several major families of eucaryotic molecular chaperones, including the Hsp60 and Hsp70 proteins. Different family members function in different organelles. Thus, as discussed in Chapter 12, mitochondria contain their own Hsp60 and Hsp70 molecules that are distinct from those that function in the cytosol; and a special Hsp70 (called *BIP*) helps to fold proteins in the endoplasmic reticulum.



(A) (B)

Figure 6–83 The structure of a molten globule. (A) A molten globule form of cytochrome b_{562} is more open and less highly ordered than the final folded form of the protein, shown in (B). Note that the molten globule contains most of the secondary structure of the final form, although the ends of the α helices are unravelled and one of the helices is only partly formed. (Courtesy of Joshua Wand, from Y. Feng et al., *Nat. Struct. Biol.* 1:30–35, 1994. With permission from Macmillan Publishers Ltd.)

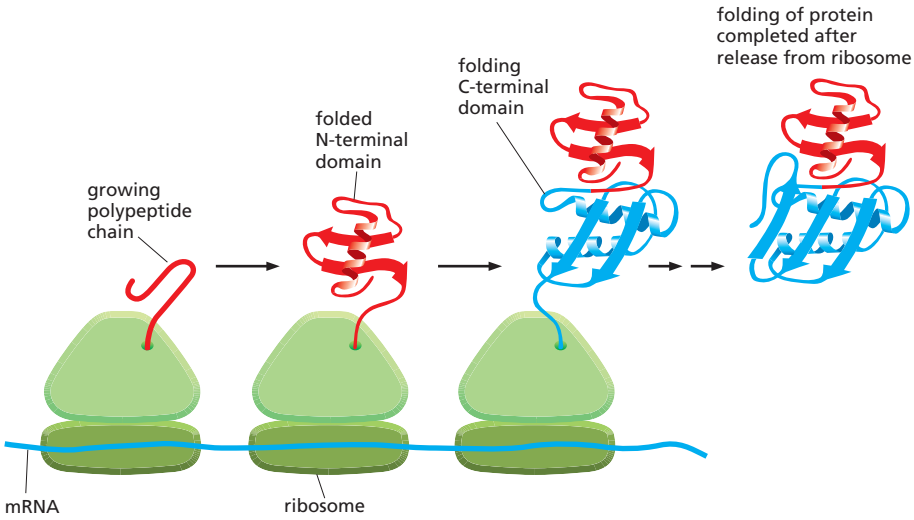


Figure 6–84 Co-translational protein folding. A growing polypeptide chain is shown acquiring its secondary and tertiary structure as it emerges from a ribosome. The N-terminal domain folds first, while the C-terminal domain is still being synthesized. This protein has not achieved its final conformation at the time it is released from the ribosome. (Modified from A.N. Federov and T.O. Baldwin, *J. Biol. Chem.* 272:32715–32718, 1997.)

The Hsp60 and Hsp70 proteins each work with their own small set of associated proteins when they help other proteins to fold. Hsps share an affinity for the exposed hydrophobic patches on incompletely folded proteins, and they hydrolyze ATP, often binding and releasing their protein substrate with each cycle of ATP hydrolysis. In other respects, the two types of Hsp proteins function differently. The Hsp70 machinery acts early in the life of many proteins, binding to a string of about seven hydrophobic amino acids before the protein leaves the ribosome (**Figure 6–86**). In contrast, Hsp60-like proteins form a large barrel-shaped structure that acts after a protein has been fully synthesized. This type of chaperone, sometimes called a *chaperonin*, forms an "isolation chamber" into which misfolded proteins are fed, preventing their aggregation and providing them with a favorable environment in which to attempt to refold (**Figure 6–87**).

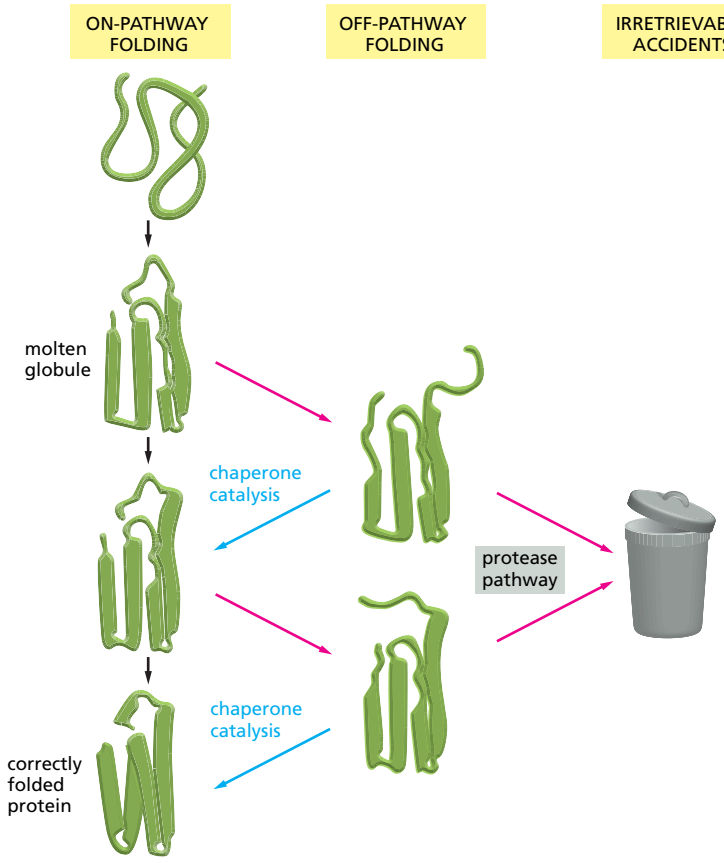


Figure 6–85 A current view of protein folding. Each domain of a newly synthesized protein rapidly attains a "molten globule" state. Subsequent folding occurs more slowly and by multiple pathways, often involving the help of a molecular chaperone. Some molecules may still fail to fold correctly; as explained in the text, specific proteases recognize and degrade these molecules.

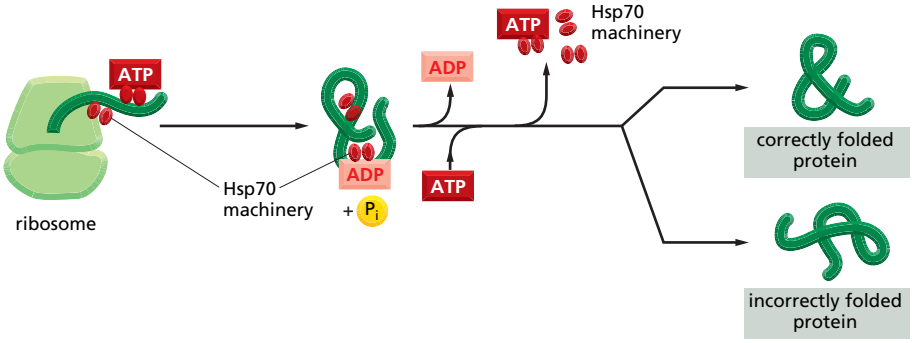


Figure 6–86 The Hsp70 family of molecular chaperones. These proteins act early, recognizing a small stretch of hydrophobic amino acids on a protein's surface. Aided by a set of smaller Hsp40 proteins (not shown), ATP-bound Hsp70 molecules grasp their target protein and then hydrolyze ATP to ADP, undergoing conformational changes that cause the Hsp70 molecules to associate even more tightly with the target. After the Hsp40 dissociates, the rapid rebinding of ATP induces the dissociation of the Hsp70 protein after ADP release. In reality, repeated cycles of Hsp protein binding and release help the target protein to refold, as schematically illustrated in Figure 6–85.

The chaperones shown in Figures 6–86 and 6–87 often use many cycles of ATP hydrolysis to fold a single polypeptide chain correctly. Although some of this energy expenditure is used to perform mechanical work, probably much more is expended to ensure that protein folding is accurate. Just as we saw for transcription, splicing, and translation, the expenditure of free energy can be used by cells to improve the accuracy of a biological process. In the case of protein folding, ATP hydrolysis allows chaperones to recognize a wide variety of misfolded structures, to halt any further misfolding and to recommence folding of a protein in an orderly way.

Although our discussion focuses on only two types of chaperones, the cell has a variety of others. The enormous diversity of proteins in cells presumably requires a wide range of chaperones with versatile surveillance and correction capabilities.

Exposed Hydrophobic Regions Provide Critical Signals for Protein Quality Control

If radioactive amino acids are added to cells for a brief period, the newly synthesized proteins can be followed as they mature into their final functional form.

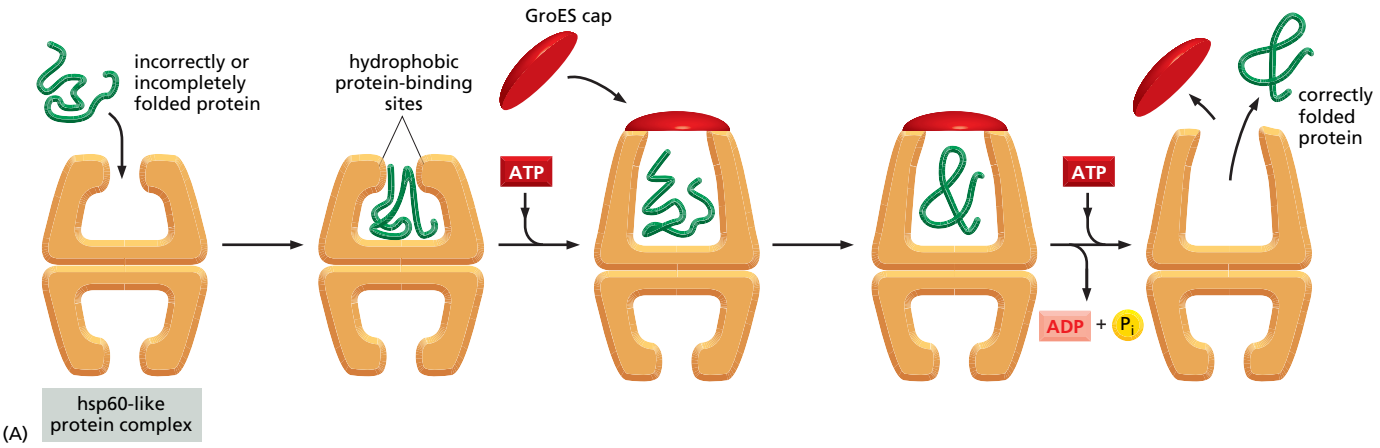
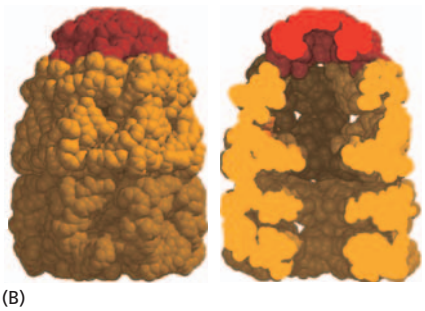


Figure 6–87 The structure and function of the Hsp60 family of molecular chaperones. (A) The catalysis of protein refolding. A misfolded protein is initially captured by hydrophobic interactions along one rim of the barrel. The subsequent binding of ATP plus a protein cap increases the diameter of the barrel rim, which may transiently stretch (partly unfold) the client protein. This also confines the protein in an enclosed space, where it has a new opportunity to fold. After about 15 seconds, ATP hydrolysis occurs, weakening the complex. Subsequent binding of another ATP molecule ejects the protein, whether folded or not, and the cycle repeats. This type of molecular chaperone is also known as a chaperonin; it is designated as Hsp60 in mitochondria, TCP1 in the cytosol of vertebrate cells, and GroEL in bacteria. As indicated, only half of the symmetrical barrel operates on a client protein at any one time. (B) The structure of GroEL bound to its GroES cap, as determined by X-ray crystallography. On the *left* is shown the outside of the barrel-like structure and on the *right* a cross section through its center. (B, adapted from B. Bukau and A.L. Horwich, *Cell* 92:351–366, 1998. With permission from Elsevier.)



This type of experiment demonstrates that the Hsp70 proteins act first, beginning when a protein is still being synthesized on a ribosome, and the Hsp60-like proteins act only later to help fold completed proteins. But how does the cell distinguish misfolded proteins, which require additional rounds of ATP-catalyzed refolding, from those with correct structures?

Before answering, we need to pause to consider the post-translational fate of proteins more broadly. Usually, if a protein has a sizable exposed patch of hydrophobic amino acids on its surface, it is abnormal: it has either failed to fold correctly after leaving the ribosome, suffered an accident that partly unfolded it at a later time, or failed to find its normal partner subunit in a larger protein complex. Such a protein is not merely useless to the cell, it can be dangerous. Many proteins with an abnormally exposed hydrophobic region can form large aggregates in the cell. We shall see that, in rare cases, such aggregates do form and cause severe human diseases. Normally, however, powerful protein quality control mechanisms prevent such disasters.

Given this background, it is not surprising that cells have evolved elaborate mechanisms that recognize the hydrophobic patches on proteins and minimize the damage they cause. Two of these mechanisms depend on the molecular chaperones just discussed, which bind to the patch and attempt to repair the defective protein by giving it another chance to fold. At the same time, by covering the hydrophobic patches, these chaperones transiently prevent protein aggregation. Proteins that very rapidly fold correctly on their own do not display such patches and the chaperones bypass them.

Figure 6–88 outlines all of the quality control choices that a cell makes for a difficult-to-fold, newly synthesized protein. As indicated, when attempts to refold a protein fail, a third mechanism is called into play that completely destroys the protein by proteolysis. The proteolytic pathway begins with the recognition of an abnormal hydrophobic patch on a protein's surface, and it ends with the delivery of the entire protein to a protein destruction machine, a complex protease known as the *proteasome*. As described next, this process depends on an elaborate protein-marking system that also carries out other central functions in the cell by destroying selected normal proteins.

The Proteasome Is a Compartmentalized Protease with Sequestered Active Sites

The proteolytic machinery and the chaperones compete with one another to reorganize a misfolded protein. If a newly synthesized protein folds rapidly, at most only a small fraction of it is degraded. In contrast, a slowly folding protein is vulnerable to the proteolytic machinery for a longer time, and many more of its molecules are destroyed before the remainder attain the proper folded state. Due to mutations or to errors in transcription, RNA splicing, and translation, some proteins never fold properly. It is particularly important that the cell destroy these potentially harmful proteins.

The apparatus that deliberately destroys aberrant proteins is the **proteasome**, an abundant ATP-dependent protease that constitutes nearly 1% of cell protein. Present in many copies dispersed throughout the cytosol and the nucleus, the proteasome also destroys aberrant proteins of the endoplasmic

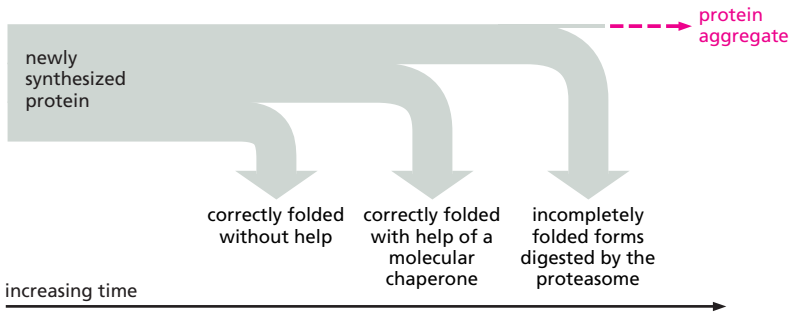


Figure 6–88 The processes that monitor protein quality following protein synthesis. A newly synthesized protein sometimes folds correctly and assembles on its own with its partner proteins, in which case the quality control mechanisms leave it alone. Incompletely folded proteins are helped to refold by molecular chaperones: first by a family of Hsp70 proteins, and then in some cases, by Hsp60-like proteins. For both types of chaperones, the client proteins are recognized by an abnormally exposed patch of hydrophobic amino acids on their surface. These "protein-rescue" processes compete with another mechanism that, upon recognizing an abnormally exposed patch, marks the protein for destruction by the proteasome. The combined activity of all of these processes is needed to prevent massive protein aggregation in a cell, which can occur when many hydrophobic regions on proteins clump together nonspecifically.

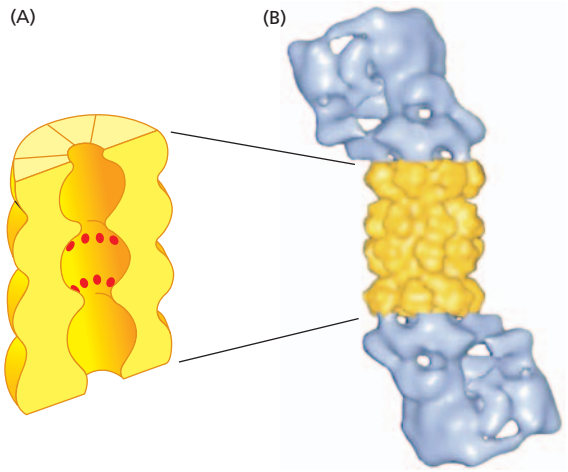


Figure 6–89 The proteasome. (A) A cutaway view of the structure of the central 20S cylinder, as determined by x-ray crystallography, with the active sites of the proteases indicated by *red dots*. (B) The entire proteasome, in which the central cylinder *(yellow)* is supplemented by a 19S cap *(blue)* at each end. The cap structure has been determined by computer processing of electron microscope images. The complex cap (also called the regulatory particle) selectively binds proteins that have been marked by ubiquitin for destruction; it then uses ATP hydrolysis to unfold their polypeptide chains and feed them through a narrow channel (see Figure 6–91) into the inner chamber of the 20S cylinder for digestion to short peptides. (B, from W. Baumeister et al., *Cell* 92:367–380, 1998. With permission from Elsevier.)

reticulum (ER). An ER-based surveillance system detects proteins that fail either to fold or to be assembled properly after they enter the ER, and *retrotranslocates* them back to the cytosol for degradation (discussed in Chapter 12).

Each proteasome consists of a central hollow cylinder (the 20S core proteasome) formed from multiple protein subunits that assemble as a quasi-cylindrical stack of four heptameric rings (**Figure 6–89**). Some of the subunits are distinct proteases whose active sites face the cylinder's inner chamber. The design prevents these highly efficient proteases from running rampant through the cell. Each end of the cylinder is normally associated with a large protein complex (the 19S cap), which contains a six-subunit protein ring, through which target proteins are threaded into the proteasome core where they are degraded (**Figure 6–90**). The threading reaction, driven by ATP hydrolysis, unfolds the target proteins as they move through the cap, exposing them to the proteases lining the proteasome core (**Figure 6–91**). The proteins that make up the ring structure in the proteasome cap belong to a large class of protein "unfoldases" known as *AAA proteins*. Many of them function as hexamers, and it is possible that they share mechanistic features with the ATP-dependent unwinding of DNA by DNA helicases (see Figure 5–15).

A crucial property of the proteasome, and one reason for the complexity of its design, is the *processivity* of its mechanism: in contrast to a "simple" protease that cleaves a substrate's polypeptide chain just once before dissociating, the proteasome keeps the entire substrate bound until all of it is converted into short peptides.

The 19S caps also act as regulated "gates" at the entrances to the inner proteolytic chamber, and they are responsible for binding a targeted protein substrate to the proteasome. With a few exceptions, the proteasomes act on proteins that have been specifically marked for destruction by the covalent attachment of a recognition tag formed from a small protein called *ubiquitin* (**Figure 6–92**A). Ubiquitin exists in cells either free or covalently linked to

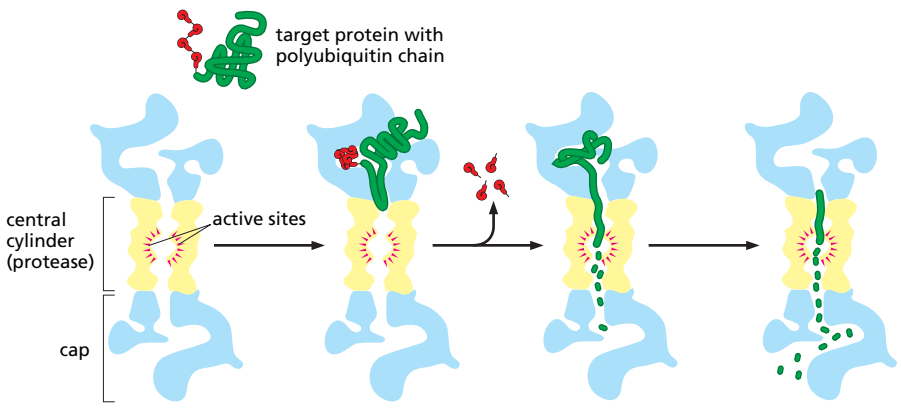


Figure 6–90 Processive protein digestion by the proteasome. The proteasome cap recognizes a substrate protein, in this case marked by a polyubiquitin chain (see Figure 6–92), and subsequently translocates it into the proteasome core, where it is digested. At an early stage, the ubiquitin is cleaved from the substrate protein and is recycled. Translocation into the core of the proteasome is mediated by a ring of ATP-dependent proteins that unfold the substrate protein as it is threaded through the ring and into the proteasome core (see Figure 6–91). (From S. Prakash and A. Matouschek, *Trends Biochem. Sci.* 29:593–600, 2004. With permission from Elsevier.)

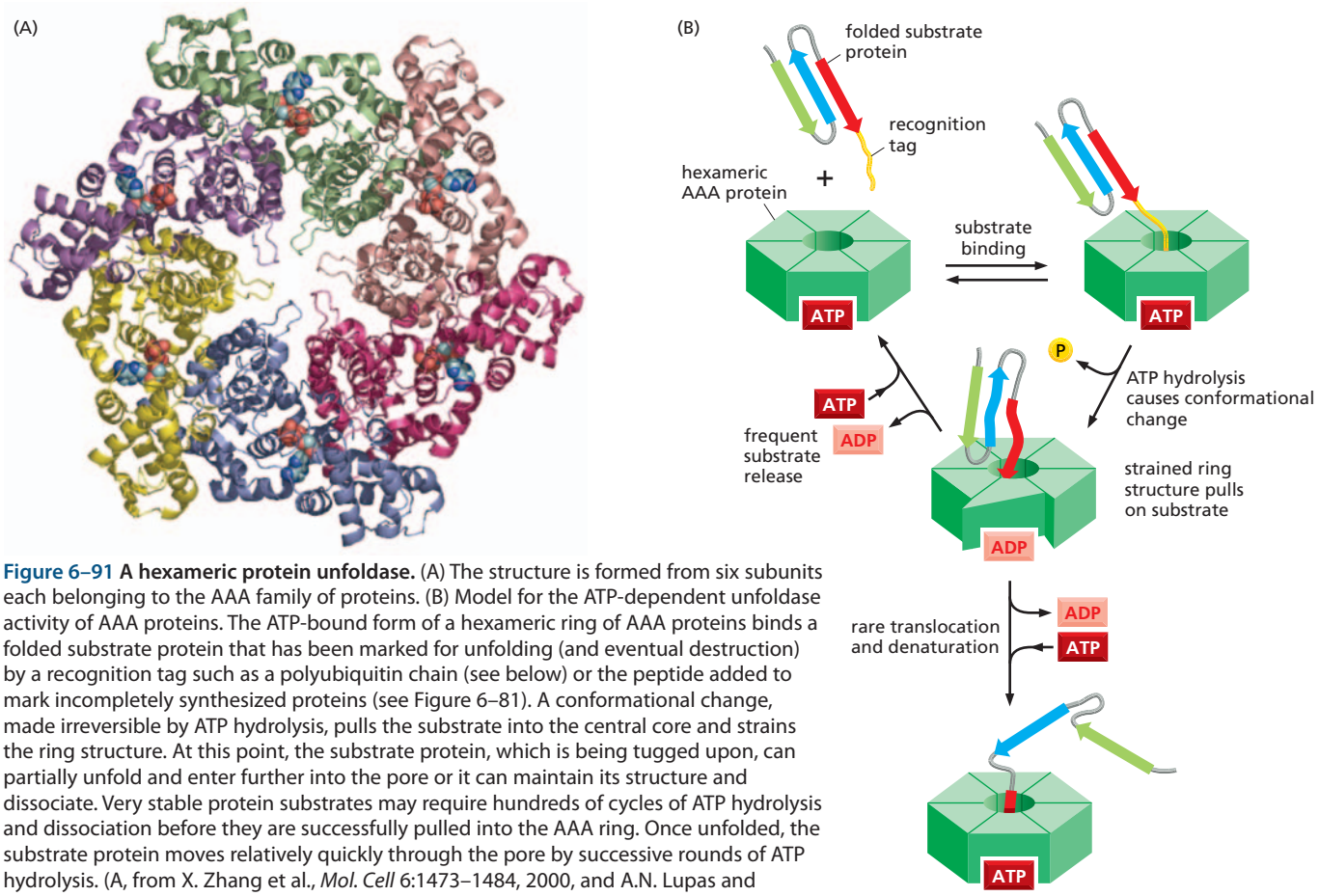


Figure 6–91 A hexameric protein unfoldase. (A) The structure is formed from six subunits each belonging to the AAA family of proteins. (B) Model for the ATP-dependent unfoldase activity of AAA proteins. The ATP-bound form of a hexameric ring of AAA proteins binds a folded substrate protein that has been marked for unfolding (and eventual destruction) by a recognition tag such as a polyubiquitin chain (see below) or the peptide added to mark incompletely synthesized proteins (see Figure 6–81). A conformational change, made irreversible by ATP hydrolysis, pulls the substrate into the central core and strains the ring structure. At this point, the substrate protein, which is being tugged upon, can partially unfold and enter further into the pore or it can maintain its structure and dissociate. Very stable protein substrates may require hundreds of cycles of ATP hydrolysis and dissociation before they are successfully pulled into the AAA ring. Once unfolded, the substrate protein moves relatively quickly through the pore by successive rounds of ATP hydrolysis. (A, from X. Zhang et al., *Mol. Cell* 6:1473–1484, 2000, and A.N. Lupas and J. Martin, *Curr. Opin. Struct. Biol.* 12:746–753, 2002; B, from R.T. Sauer et al., *Cell* 119:9–18, 2004. All with permission from Elsevier.)

many different intracellular proteins. For many proteins, tagging by ubiquitin results in their destruction by the proteasome. However, in other cases, ubiquitin tagging has an entirely different meaning. Ultimately, it is the number of ubiquitin molecules added and the way in which they are linked together that determines how the cell interprets the ubiquitin message (**Figure 6–93**). In the following sections, we emphasize the role of ubiquitylation in signifying protein degradation.

An Elaborate Ubiquitin-Conjugating System Marks Proteins for Destruction

Ubiquitin is prepared for conjugation to other proteins by the ATP-dependent *ubiquitin-activating enzyme* (E1), which creates an activated, E1-bound ubiquitin that is subsequently transferred to one of a set of *ubiquitin-conjugating* (E2) enzymes (Figure 6–92B). The E2 enzymes act in conjunction with accessory (E3) proteins. In the E2–E3 complex, called *ubiquitin ligase*, the E3 component binds to specific degradation signals, called degrons, in protein substrates, helping E2 to form a *polyubiquitin* chain linked to a lysine of the substrate protein. In this chain, the C-terminal residue of each ubiquitin is linked to a specific lysine of the preceding ubiquitin molecule (see Figure 6–93), producing a linear series of ubiquitin–ubiquitin conjugates (Figure 6–92C). It is this polyubiquitin chain on a target protein that is recognized by a specific receptor in the proteasome.

There are roughly 30 structurally similar but distinct E2 enzymes in mammals, and hundreds of different E3 proteins that form complexes with specific E2 enzymes. The ubiquitin–proteasome system thus consists of many distinct but similarly organized proteolytic pathways, which have in common both the

E1 enzyme at the "top" and the proteasome at the "bottom," and differ by the compositions of their E2–E3 ubiquitin ligases and accessory factors. Distinct ubiquitin ligases recognize different degradation signals, and therefore target distinct subsets of intracellular proteins for destruction.

Denatured or otherwise misfolded proteins, as well as proteins containing oxidized or other abnormal amino acids, are recognized and destroyed because abnormal proteins tend to present on their surface amino acid sequences or conformational motifs that are recognized as degradation signals by a set of E3 molecules in the ubiquitin–proteasome system; these sequences must of course be buried and therefore inaccessible in the normal counterparts of these proteins. However, a proteolytic pathway that recognizes and destroys abnormal proteins must be able to distinguish between *completed* proteins that have "wrong" conformations and the many growing polypeptides on ribosomes (as well as polypeptides just released from ribosomes) that have not yet achieved their normal folded conformation. This is not a trivial problem; the ubiquitin–proteasome system is thought to destroy many of the nascent and newly formed protein molecules not because these proteins are abnormal as such, but because they transiently expose degradation signals that are buried in their mature (folded) state.

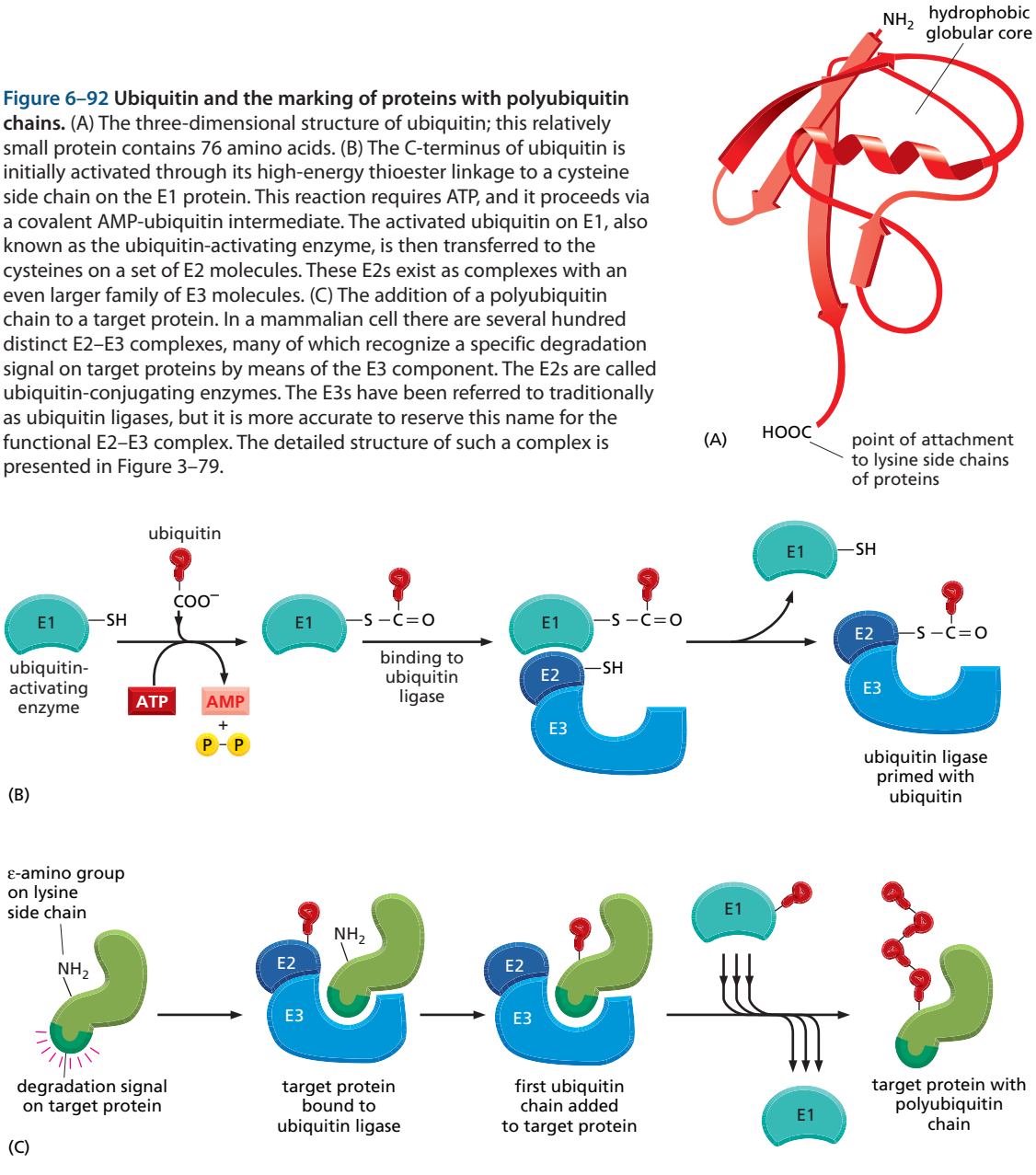


Figure 6–92 Ubiquitin and the marking of proteins with polyubiquitin chains. (A) The three-dimensional structure of ubiquitin; this relatively small protein contains 76 amino acids. (B) The C-terminus of ubiquitin is initially activated through its high-energy thioester linkage to a cysteine side chain on the E1 protein. This reaction requires ATP, and it proceeds via a covalent AMP-ubiquitin intermediate. The activated ubiquitin on E1, also known as the ubiquitin-activating enzyme, is then transferred to the cysteines on a set of E2 molecules. These E2s exist as complexes with an even larger family of E3 molecules. (C) The addition of a polyubiquitin chain to a target protein. In a mammalian cell there are several hundred distinct E2–E3 complexes, many of which recognize a specific degradation signal on target proteins by means of the E3 component. The E2s are called ubiquitin-conjugating enzymes. The E3s have been referred to traditionally as ubiquitin ligases, but it is more accurate to reserve this name for the functional E2–E3 complex. The detailed structure of such a complex is presented in Figure 3–79.

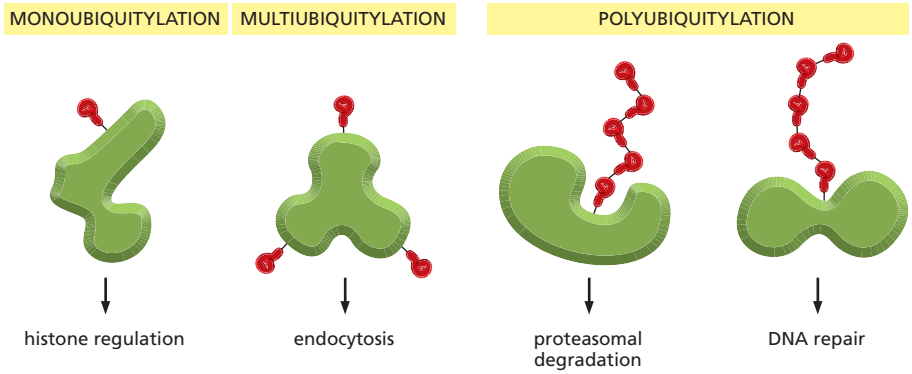


Figure 6–93 The marking of proteins by ubiquitin. Each modification pattern shown can have a specific meaning to the cell. The two types of polyubiquitylation differ in the way the ubiquitin molecules are linked together. Linkage through Lys48 signifies degradation by the proteasome whereas that through Lys63 has other meanings. Ubiquitin markings are "read" by proteins that specifically recognize each type of modification.

Many Proteins Are Controlled by Regulated Destruction

One function of intracellular proteolytic mechanisms is to recognize and eliminate misfolded or otherwise abnormal proteins, as just described. Yet another function of these proteolytic pathways is to confer short lifetimes on specific normal proteins whose concentrations must change promptly with alterations in the state of a cell. Some of these short-lived proteins are degraded rapidly at all times, while many others are *conditionally* short-lived, that is, they are metabolically stable under some conditions but become unstable upon a change in the cell's state. For example, mitotic cyclins are long-lived throughout the cell cycle until their sudden degradation at the end of mitosis, as explained in Chapter 17.

How is such a regulated destruction of a protein controlled? Several mechanisms are illustrated through specific examples that appear later in this book. In one general class of mechanism (**Figure 6–94**A), the activity of a ubiquitin ligase is turned on either by E3 phosphorylation or by an allosteric transition in an E3 protein caused by its binding to a specific small or large molecule. For example, the anaphase-promoting complex (APC) is a multisubunit ubiquitin ligase that is activated by a cell-cycle-timed subunit addition at mitosis. The activated APC then causes the degradation of mitotic cyclins and several other regulators of the metaphase–anaphase transition (see Figure 17–44).

Alternatively, in response either to intracellular signals or to signals from the environment, a degradation signal can be created in a protein, causing its rapid ubiquitylation and destruction by the proteasome. One common way to create such a signal is to phosphorylate a specific site on a protein that unmasks a normally hidden degradation signal. Another way to unmask such a signal is by the regulated dissociation of a protein subunit. Finally, powerful degradation signals can be created by cleaving a single peptide bond, provided that this cleavage creates a new N-terminus that is recognized by a specific E3 as a "destabilizing" N-terminal residue (Figure 6–94B).

The N-terminal type of degradation signal arises because of the "N-end rule," which relates the lifetime of a protein *in vivo* to the identity of its N-terminal residue. There are 12 destabilizing residues in the N-end rule of the yeast *S. cerevisiae* (Arg, Lys, His, Phe, Leu, Tyr, Trp, Ile, Asp, Glu, Asn, and Gln), out of the 20 standard amino acids. The destabilizing N-terminal residues are recognized by a special ubiquitin ligase that is conserved from yeast to humans.

As we have seen, all proteins are initially synthesized bearing methionine (or formylmethionine in bacteria), as their N-terminal residue, which is a stabilizing residue in the N-end rule. Special proteases, called methionine aminopeptidases, will often remove the first methionine of a nascent protein, but they will do so only if the second residue is also stabilizing according to N-end rule. Therefore, it was initially unclear how N-end rule substrates form *in vivo*. However, it is now understood that these substrates are formed by site-specific proteases. For example, a subunit of cohesin, a protein complex that holds sister chromatids together, is cleaved by a highly specific protease during the metaphase–anaphase transition. This cell-cycle-regulated cleavage allows separation of the sister chromatids and leads to the completion of mitosis (see

Figure 17–44). The C-terminal fragment of the cleaved subunit bears an N-terminal arginine, a destabilizing residue in the N-end rule. Mutant cells lacking the N-end rule pathway exhibit a greatly increased frequency of chromosome loss, presumably because a failure to degrade this fragment of the cohesin subunit interferes with the formation of new chromatid-associated cohesin complexes in the next cell cycle.

Abnormally Folded Proteins Can Aggregate to Cause Destructive Human Diseases

Many inherited human diseases (for example, sickle-cell anemia (see p. 1495) and α-1-antitrypsin deficiency, a condition that often leads to liver disease and emphysema) result from mutant proteins that escape the cell's quality controls, fold abnormally, and form aggregates. By absorbing critical macromolecules, these aggregates can severely damage cells and even cause cell death. Often, the inheritance of a single mutant allele of a gene can cause disease, since the normal copy of the gene cannot protect the cell from the destructive properties of the aggregate.

In normal humans, the gradual decline of the cell's protein quality controls can also cause disease by permitting normal proteins to form aggregates (**Figure 6–95**). In some cases, the protein aggregates are released from dead cells and accumulate in the extracellular matrix that surrounds the cells in a tissue, and in extreme cases they can also damage tissues. Because the brain is composed of a highly organized collection of nerve cells, it is especially vulnerable.

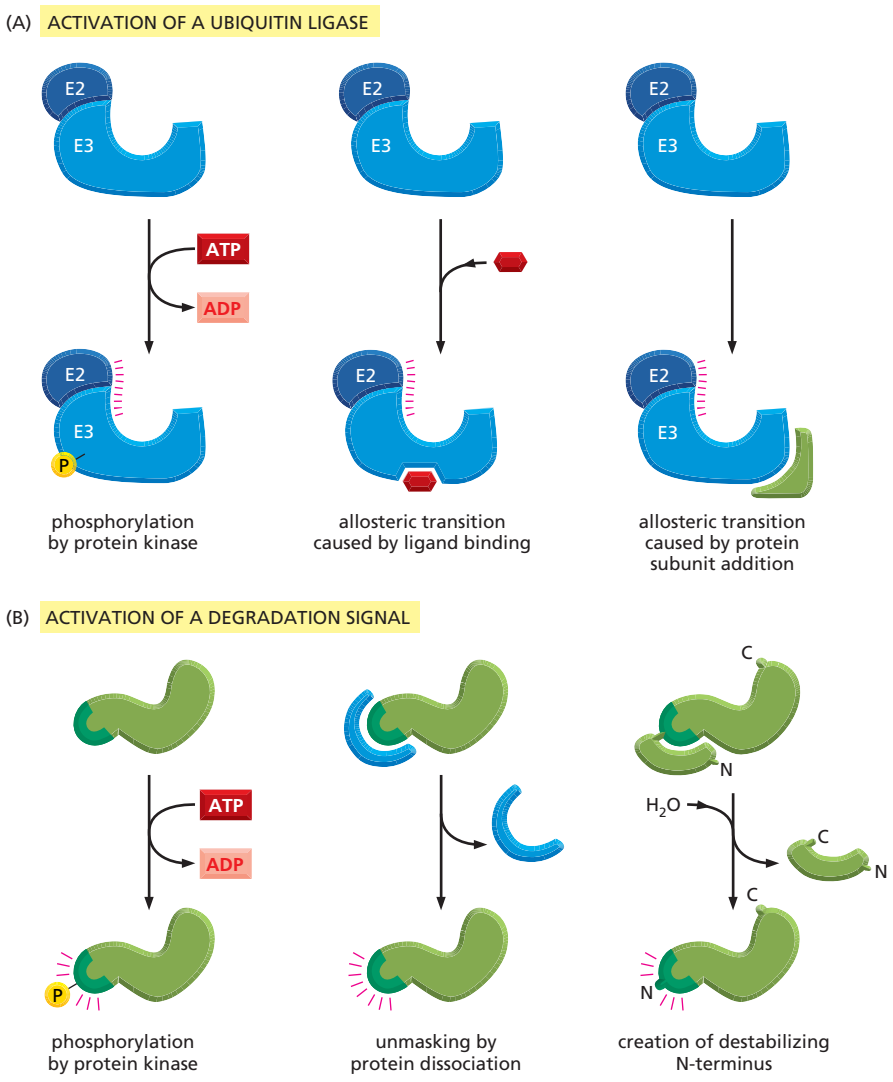


Figure 6–94 Two general ways of inducing the degradation of a specific protein. (A) Activation of a specific E3 molecule creates a new ubiquitin ligase. (B) Creation of an exposed degradation signal in the protein to be degraded. This signal binds a ubiquitin ligase, causing the addition of a polyubiquitin chain to a nearby lysine on the target protein. All six pathways shown are known to be used by cells to induce the movement of selected proteins into the proteasome.

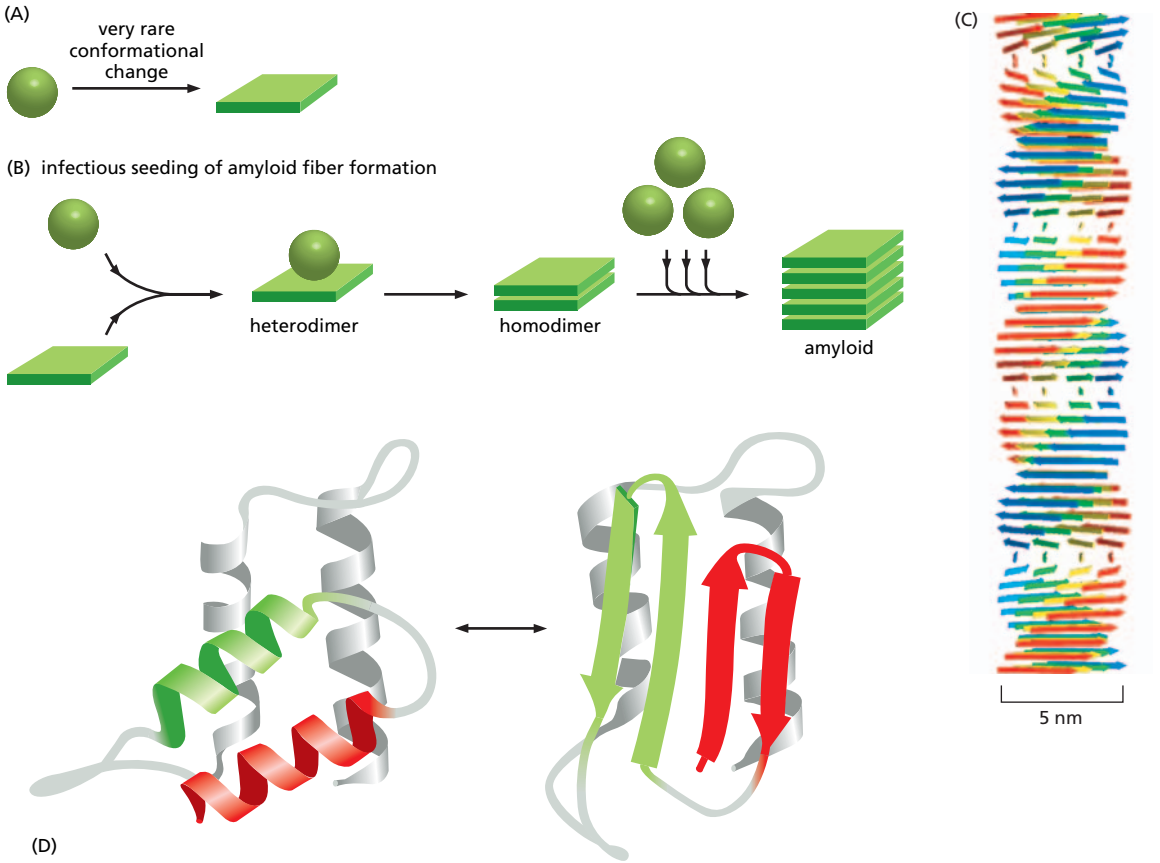


Figure 6–95 Protein aggregates that cause human disease. (A) Schematic illustration of the type of conformational change in a protein that produces material for a cross-beta filament. (B) Diagram illustrating the self-infectious nature of the protein aggregation that is central to prion diseases. PrP (prion protein) is highly unusual because the misfolded version of the protein, called PrP*, induces the normal PrP protein it contacts to change its conformation, as shown. Most of the human diseases caused by protein aggregation are caused by the overproduction of a variant protein that is especially prone to aggregation, but the protein aggregate cannot spread from one animal to another. (C) Drawing of a cross-beta filament, a common type of protease-resistant protein aggregate found in many human neurological diseases. Because the hydrogen-bond interactions in a β sheet form between polypeptide backbone atoms (see Figure 3–9), a number of different abnormally folded proteins can produce this structure. (D) One of several possible models for the conversion of PrP to PrP*, showing the likely change of two α helices into four β strands. Although the structure of the normal protein has been determined accurately, the structure of the infectious form is not yet known with certainty because the aggregation has prevented the use of standard structural techniques. (C, courtesy of Louise Serpell, adapted from M. Sunde et al., *J. Mol. Biol.* 273:729–739, 1997. With permission from Academic Press; D, adapted from S.B. Prusiner, *Trends Biochem. Sci.* 21:482–487, 1996. With permission from Elsevier.)

Not surprisingly, therefore, protein aggregates primarily cause neurodegenerative diseases. Prominent among these are Huntington's disease and Alzheimer's disease—the latter causing age-related dementia in more than 20 million people in today's world.

For a particular type of protein aggregate to survive, grow, and damage an organism, it must be highly resistant to proteolysis both inside and outside the cell. Many of the protein aggregates that cause problems form fibrils built from a series of polypeptide chains that are layered one over the other as a continuous stack of β sheets. This so-called *cross-beta filament* (Figure 6–95C), a structure particularly resistant to proteolysis, is observed in many of the neurological disorders caused by protein aggregates, where it produces distinctly staining deposits known as *amyloids*.

One particular variety of these pathologies has attained special notoriety. These are the **prion diseases**. Unlike Huntington's or Alzheimer's, prion diseases can spread from one organism to another, providing that the second organism eats a tissue containing the protein aggregate. A set of diseases—called scrapie in sheep, Creutzfeldt–Jacob disease (CJD) in humans, and bovine spongiform encephalopathy (BSE) in cattle—are caused by a misfolded, aggregated form of a protein called PrP (for prion protein). The PrP is normally located on the outer surface of the plasma membrane, most prominently in neurons. Its normal

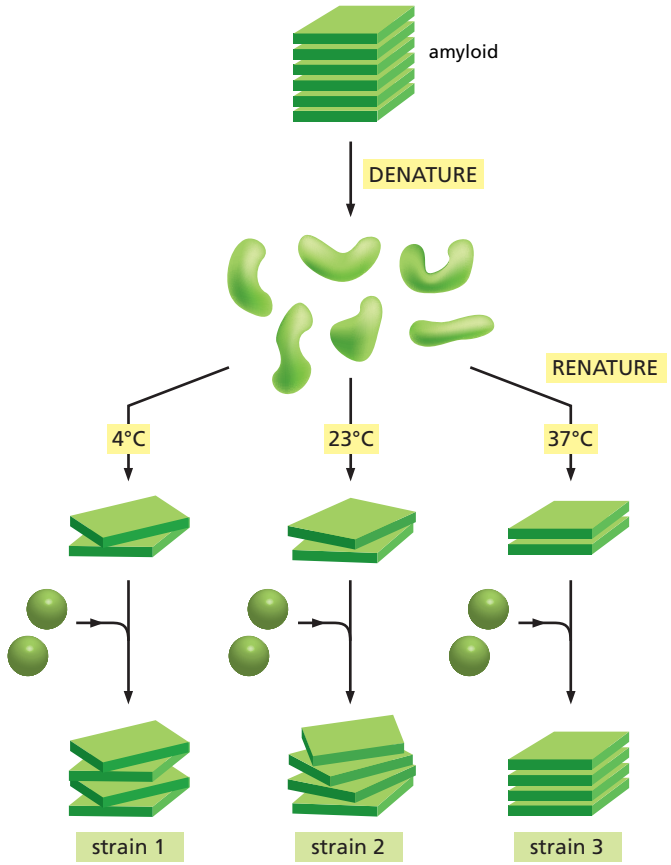


Figure 6–96 Creation of different prion strains *in vitro*. In this experiment, amyloid fibers were denatured and the components renatured at different temperatures. This treatment produced three distinctive types of amyloids, each of which could self-propagate when new subunits are added.

function is not known. However, PrP has the unfortunate property of being convertible to a very special abnormal conformation (see Figure 6–95A). This conformation not only forms protease-resistant, cross-beta filaments; it is also "infectious" because it converts normally folded molecules of PrP to the same pathological form. This property creates a positive feedback loop that propagates the abnormal form of PrP, called PrP* (see Figure 6–95B) and thereby allows the pathological conformation to spread rapidly from cell to cell in the brain, eventually causing death in both animals and humans. It can be dangerous to eat the tissues of animals that contain PrP*, as witnessed by the spread of BSE (commonly referred to as "mad cow disease") from cattle to humans in Great Britain. Fortunately, in the absence of PrP*, PrP is extraordinarily difficult to convert to its abnormal form.

Although very few proteins have the potential to misfold into an infectious conformation, another example causes an otherwise mysterious "protein-only inheritance" observed in yeast cells. The ability to study infectious proteins in yeast has clarified another remarkable feature of prions. These protein molecules can form several distinctively different types of aggregates from the same polypeptide chain. Moreover, each type of aggregate can be infectious, forcing normal protein molecules to adopt the same type of abnormal structure. Thus, several different "strains" of infectious particles can arise from the same polypeptide chain (**Figure 6–96**). How a single polypeptide sequence can adopt multiple aggregate forms is not fully understood; it is possible that all prion aggregates resemble cross-beta filaments (see Figure 6–95C) where the structure is held together predominantly with main peptide chain interactions. This would leave the amino acid side chains free to adopt different conformations and, if the structures are self-propagating, the existence of different strains could be explained.

Finally, although prions were discovered because they cause disease, they also appear to have some positive roles in the cell. For example, some species of fungi use prion transformations to establish different types of cells. Although the idea is controversial, it has even been proposed that prions have a role in consolidating memories in complex, multicellular organisms like ourselves.

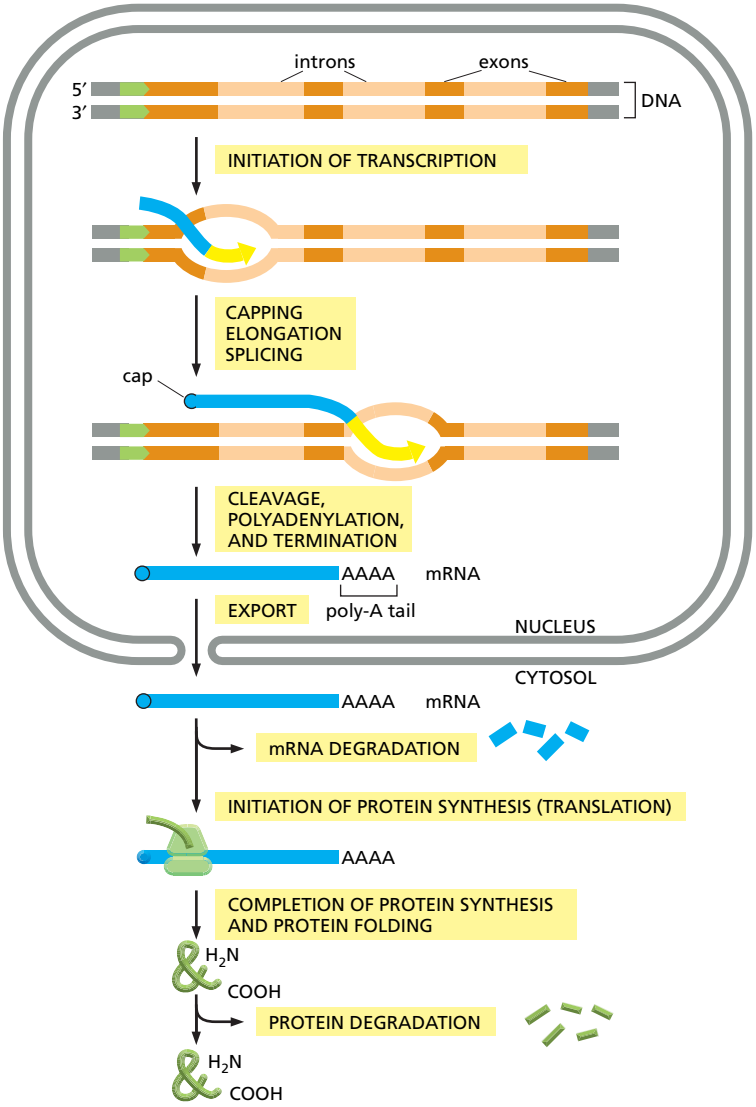


Figure 6–97 The production of a protein by a eucaryotic cell. The final level of each protein in a eucaryotic cell depends upon the efficiency of each step depicted.

There Are Many Steps From DNA to Protein

We have seen so far in this chapter that many different types of chemical reactions are required to produce a properly folded protein from the information contained in a gene (**Figure 6–97**). The final level of a properly folded protein in a cell therefore depends upon the efficiency with which each of the many steps is performed.

In the following chapter, we shall see that cells have the ability to change the levels of their proteins according to their needs. In principle, any or all of the steps in Figure 6–97 could be regulated for each individual protein. As we shall see in Chapter 7, there are examples of regulation at each step from gene to protein. However, the initiation of transcription is the most common point for a cell to regulate the expression of each of its genes. This makes sense, inasmuch as the most efficient way to keep a gene from being expressed is to block the very first step—the transcription of its DNA sequence into an RNA molecule.

Summary

The translation of the nucleotide sequence of an mRNA molecule into protein takes place in the cytoplasm on a large ribonucleoprotein assembly called a ribosome. The amino acids used for protein synthesis are first attached to a family of tRNA molecules, each of which recognizes, by complementary base-pair interactions, particular sets of three nucleotides in the mRNA (codons). The sequence of nucleotides in the mRNA is then read from one end to the other in sets of three according to the genetic code.

To initiate translation, a small ribosomal subunit binds to the mRNA molecule at a start codon (AUG) that is recognized by a unique initiator tRNA molecule. A large ribosomal subunit binds to complete the ribosome and begin protein synthesis. During this phase, aminoacyl-tRNAs—each bearing a specific amino acid—bind sequentially to the appropriate codons in mRNA through complementary base pairing between tRNA anticodons and mRNA codons. Each amino acid is added to the C-terminal end of the growing polypeptide in four sequential steps: aminoacyl-tRNA binding, followed by peptide bond formation, followed by two ribosome translocation steps. Elongation factors use GTP hydrolysis to drive these reactions forward and to improve the accuracy of amino acid selection. The mRNA molecule progresses codon by codon through the ribosome in the 5′-to-3′ direction until it reaches one of three stop codons. A release factor then binds to the ribosome, terminating translation and releasing the completed polypeptide.

Eucaryotic and bacterial ribosomes are closely related, despite differences in the number and size of their rRNA and protein components. The rRNA has the dominant role in translation, determining the overall structure of the ribosome, forming the binding sites for the tRNAs, matching the tRNAs to codons in the mRNA, and creating the active site of the peptidyl transferase enzyme that links amino acids together during translation.

In the final steps of protein synthesis, two distinct types of molecular chaperones guide the folding of polypeptide chains. These chaperones, known as Hsp60 and Hsp70, recognize exposed hydrophobic patches on proteins and serve to prevent the protein aggregation that would otherwise compete with the folding of newly synthesized proteins into their correct three-dimensional conformations. This protein folding process must also compete with an elaborate quality control mechanism that destroys proteins with abnormally exposed hydrophobic patches. In this case, ubiquitin is covalently added to a misfolded protein by a ubiquitin ligase, and the resulting polyubiquitin chain is recognized by the cap on a proteasome that moves the entire protein to the interior of the proteasome for proteolytic degradation. A closely related proteolytic mechanism, based on special degradation signals recognized by ubiquitin ligases, is used to determine the lifetimes of many normally folded proteins. By this method, selected normal proteins are removed from the cell in response to specific signals.

THE RNA WORLD AND THE ORIGINS OF LIFE

We have seen that the expression of hereditary information requires extraordinarily complex machinery and proceeds from DNA to protein through an RNA intermediate. This machinery presents a central paradox: if nucleic acids are required to synthesize proteins and proteins are required, in turn, to synthesize nucleic acids, how did such a system of interdependent components ever arise? One view is that an *RNA world* existed on Earth before modern cells arose (**Figure 6–98**). According to this hypothesis, RNA both stored genetic information and catalyzed the chemical reactions in primitive cells. Only later in evolutionary time did DNA take over as the genetic material and proteins become the major catalyst and structural component of cells. If this idea is correct, then the transition out of the RNA world was never complete; as we have seen in this chapter, RNA still catalyzes several fundamental reactions in modern-day cells, which can be viewed as molecular fossils of an earlier world.

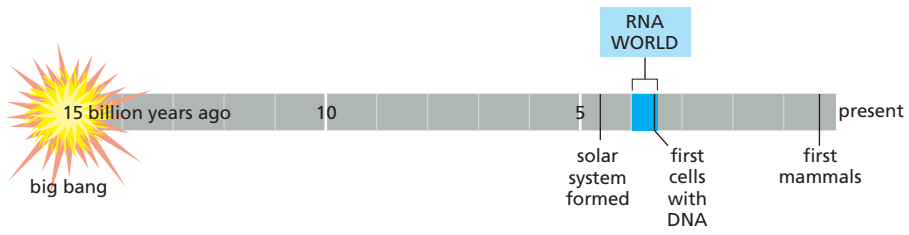


Figure 6–98 Time line for the universe, suggesting the early existence of an RNA world of living systems.

In this section we present some of the arguments in support of the RNA world hypothesis. We will see that several of the more surprising features of modern-day cells, such as the ribosome and the pre-mRNA splicing machinery, are most easily explained by viewing them as descendants of a complex network of RNA-mediated interactions that dominated cell metabolism in the RNA world. We also discuss how DNA may have taken over as the genetic material, how the genetic code may have arisen, and how proteins may have eclipsed RNA to perform the bulk of biochemical catalysis in modern-day cells.

Life Requires Stored Information

It has been proposed that the first "biological" molecules on Earth were formed by metal-based catalysis on the crystalline surfaces of minerals. In principle, an elaborate system of molecular synthesis and breakdown (metabolism) could have existed on these surfaces long before the first cells arose. Although controversial, many scientists believe that an extensive phase of "chemical evolution" took place on the prebiotic Earth, during which small molecules that could catalyze their own synthesis competed with each other for raw materials.

But life requires much more than this. As described in Chapter 1, *heredity* is perhaps the central feature of life. Not only must a cell use raw materials to create a network of catalyzed reactions, it must do so according to an elaborate set of instructions encoded in the hereditary information. The replication of this information ensures that the complex metabolism of cells can accurately reproduce itself. Another crucial feature of life is the genetic variability that results from changes in the hereditary information. This variability, acted upon by selective pressures, is responsible for the great diversity of life on our planet.

Thus, the emergence of life requires a way to store information, a way to duplicate it, a way to change it, and a way to convert the information through catalysis into favorable chemical reactions. But how could such a system begin to be formed? In present-day cells the most versatile catalysts are polypeptides, composed of many different amino acids with chemically diverse side chains and, consequently, able to adopt diverse three-dimensional forms that bristle with reactive chemical groups. Polypeptides also carry information, in the order of their amino acid subunits. But there is no known way in which a polypeptide can reproduce itself by directly specifying the formation of another of precisely the same sequence.

Polynucleotides Can Both Store Information and Catalyze Chemical Reactions

Polynucleotides have one property that contrasts with those of polypeptides: they can directly guide the formation of copies of their own sequence. This capacity depends on complementary base pairing of nucleotide subunits, which enables one polynucleotide to act as a template for the formation of another. As we have seen in this and the preceding chapter, such complementary templating mechanisms lie at the heart of DNA replication and transcription in modern-day cells.

But the efficient synthesis of polynucleotides by such complementary templating mechanisms requires catalysts to promote the polymerization reaction: without catalysts, polymer formation is slow, error-prone, and inefficient. Today, template-based nucleotide polymerization is rapidly catalyzed by protein enzymes—such as the DNA and RNA polymerases. How could such polymerization be catalyzed before proteins with the appropriate enzymatic specificity existed? The beginnings of an answer to this question came from the discovery in 1982 that RNA molecules themselves can act as catalysts. We have seen in this chapter, for example, that a molecule of RNA catalyzes one of the central reactions in the cell, the covalent joining of amino acids to form proteins. The unique potential of RNA molecules to act both as information carrier and as catalyst forms the basis of the RNA world hypothesis.

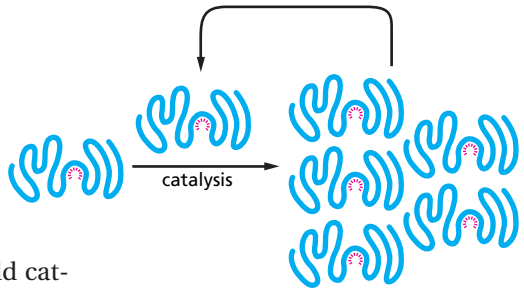


Figure 6–99 An RNA molecule that can catalyze its own synthesis. This hypothetical process would require catalysis of the production of both a second RNA strand of complementary nucleotide sequence and the use of this second RNA molecule as a template to form many molecules of RNA with the original sequence. The *red* rays represent the active site of this hypothetical RNA enzyme.

RNA therefore has all the properties required of a molecule that could catalyze a variety of chemical reactions, including those that lead to its own synthesis (**Figure 6–99**). Although self-replicating systems of RNA molecules have not been found in nature, scientists are confident that they can be constructed in the laboratory. While this demonstration would not prove that self-replicating RNA molecules were essential in the origin of life on Earth, it would certainly indicate that such a scenario is possible.

A Pre-RNA World May Predate the RNA World

Although RNA seems well suited to form the basis for a self-replicating set of biochemical catalysts, it is not clear that RNA was the first kind of molecule to do so. From a purely chemical standpoint, it is difficult to imagine how long RNA molecules could be formed initially by purely nonenzymatic means. For one thing, the precursors of RNA, the ribonucleotides, are difficult to form nonenzymatically. Moreover, the formation of RNA requires that a long series of 3′-to-5′ phosphodiester linkages assemble in the face of a set of competing reactions, including hydrolysis, 2′-to-5′ linkages, and 5′-to-5′ linkages. Given these problems, it has been suggested that the first molecules to possess both catalytic activity and information storage capabilities may have been polymers that resemble RNA but are chemically simpler (**Figure 6–100**). We do not have any remnants of these compounds in present-day cells, nor do such compounds leave fossil records. Nonetheless, the relative simplicity of these "RNA-like polymers" suggests that one of them, rather than RNA itself, may have been the first biopolymer on Earth capable of both information storage and catalytic activity.

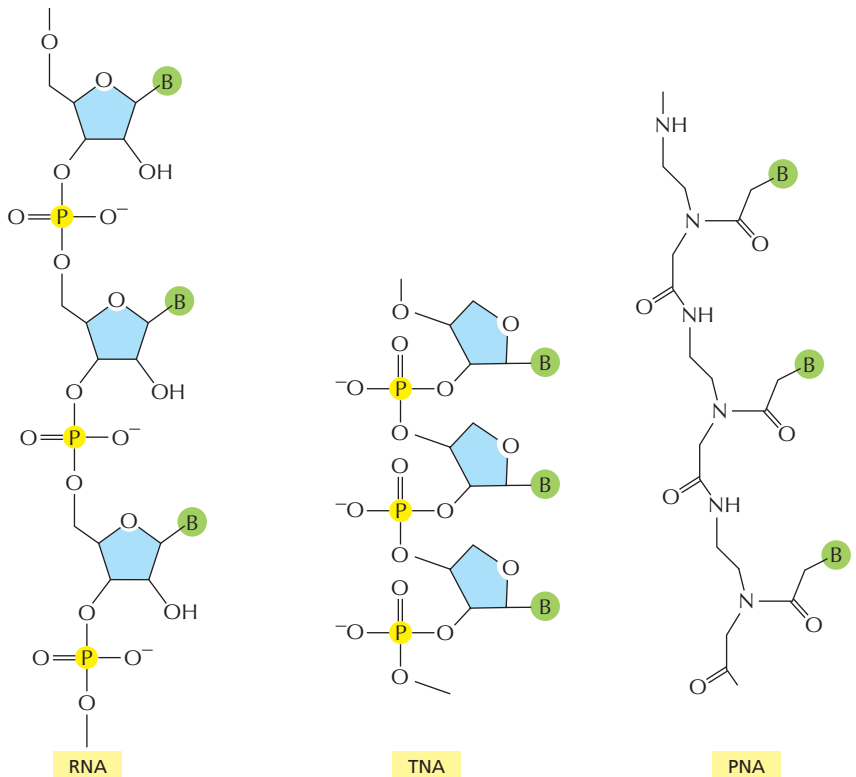


Figure 6–100 Structures of RNA and two related information-carrying polymers. In each case, B indicates a purine or pyrimidine base. The polymer TNA (threose nucleic acid) has a 4-carbon sugar unit in contrast to the 5-carbon ribose in RNA. In PNA (peptide nucleic acid), the ribose phosphate backbone of RNA has been replaced by the peptide backbone found in proteins. Like RNA, TNA and PNA can form double helices through complementary base-pairing, and each could therefore in principle serve as a template for its own synthesis.

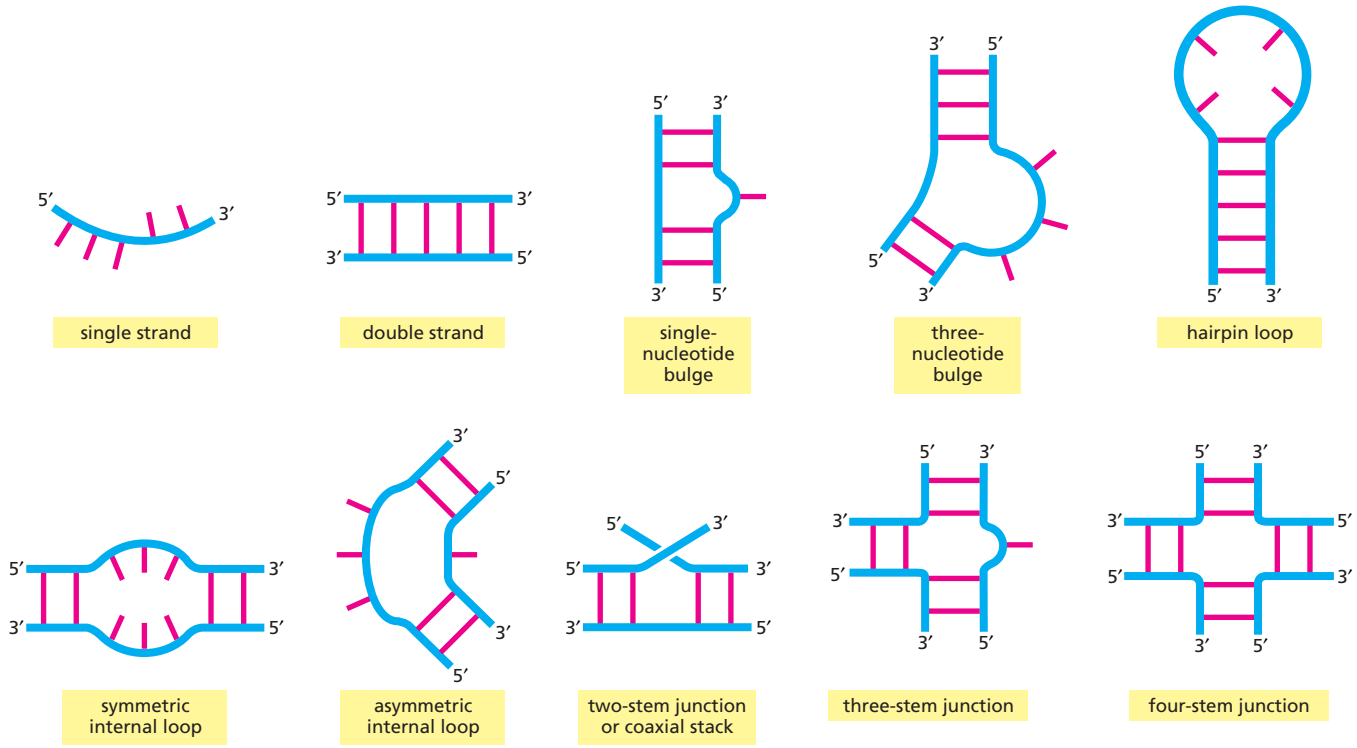


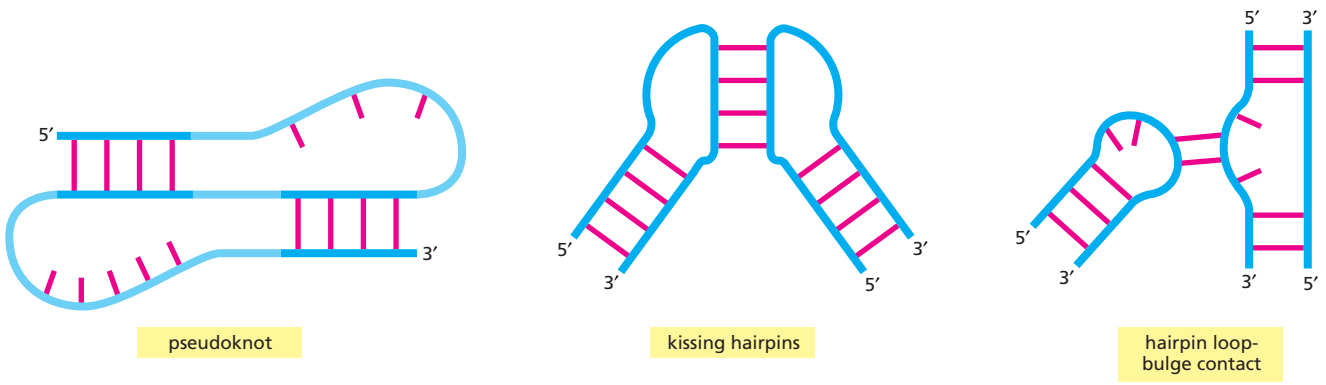
Figure 6–101 Common elements of RNA secondary structure. Conventional, complementary base-pairing interactions are indicated by *red* "rungs" in double-helical portions of the RNA.

If the pre-RNA world hypothesis is correct, then a transition to the RNA world must have occurred, presumably through the synthesis of RNA using one of these simpler polymers as both template and catalyst. While the details of the pre-RNA and RNA worlds will likely remain unknown, we know for certain that RNA molecules can catalyze a wide variety of chemical reactions, and we now turn to the properties of RNA that make this possible.

Single-Stranded RNA Molecules Can Fold into Highly Elaborate Structures

We have seen that complementary base-pairing and other types of hydrogen bonds can occur between nucleotides in the same chain, causing an RNA molecule to fold up in a unique way determined by its nucleotide sequence (see, for example, Figures 6–6, 6–52, and 6–69). Comparisons of many RNA structures have revealed conserved motifs, short structural elements that are used over and over again as parts of larger structures. **Figure 6–101** shows some of these RNA secondary structural motifs, and **Figure 6–102** shows a few common examples of more complex and often longer-range interactions, known as RNA tertiary interactions.

Figure 6–102 Examples of RNA tertiary interactions. Some of these interactions can join distant parts of the same RNA molecule or bring two separate RNA molecules together.



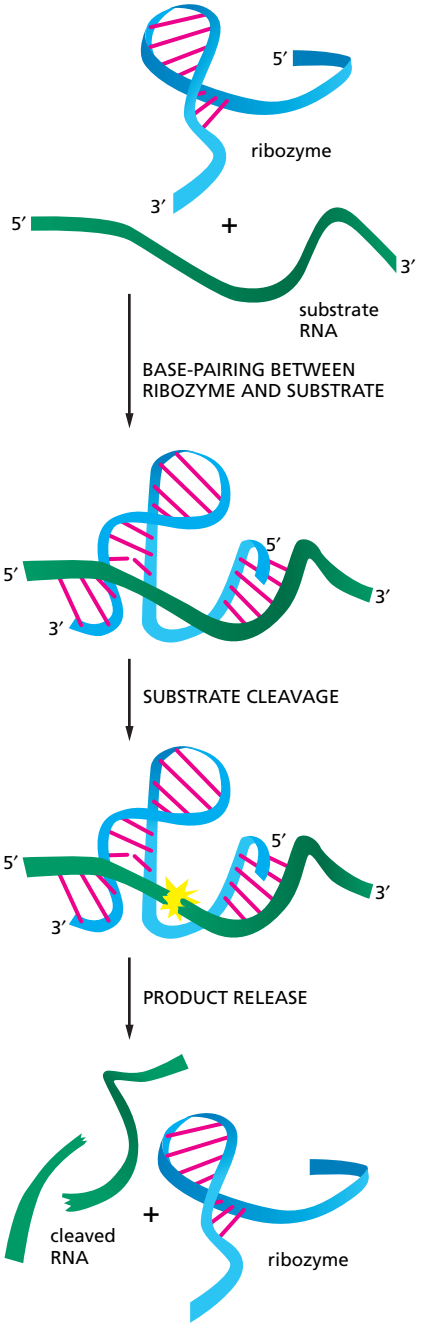


Figure 6–103 A ribozyme. This simple RNA molecule catalyzes the cleavage of a second RNA at a specific site. This ribozyme is found embedded in larger RNA genomes—called viroids—which infect plants. The cleavage, which occurs in nature at a distant location on the same RNA molecule that contains the ribozyme, is a step in the replication of the viroid genome. Although not shown in the figure, the reaction requires a Mg molecule positioned at the active site. (Adapted from T.R. Cech and O.C. Uhlenbeck, *Nature* 372:39–40, 1994. With permission from Macmillan Publishers Ltd.)

Protein catalysts require a surface with unique contours and chemical properties on which a given set of substrates can react (discussed in Chapter 3). In exactly the same way, an RNA molecule with an appropriately folded shape can serve as an enzyme (**Figure 6–103**). Like some proteins, many of these ribozymes work by positioning metal ions at their active sites. This feature gives them a wider range of catalytic activities than the limited chemical groups of the polynucleotide chain.

Relatively few catalytic RNAs are known to exist in modern-day cells, however, and much of our inference about the RNA world has come from experiments in which large pools of RNA molecules of random nucleotide sequences are generated in the laboratory. Those rare RNA molecules with a property specified by the experimenter are then selected out and studied (**Figure 6–104**). Such experiments have created RNAs that can catalyze a wide variety of biochemical reactions (**Table 6–5**), with reaction rate enhancements approaching those of proteins. Given these findings, it is not clear why protein catalysts greatly outnumber ribozymes in modern cells. Experiments have shown that RNA molecules may have more difficulty than proteins in binding to flexible, hydrophobic substrates; moreover, the availability of 20 types of amino acids over four types of bases may provide proteins with a greater number of catalytic strategies.

Like proteins, RNAs can undergo conformational changes, either in response to small molecules or to other RNAs. We saw several examples of this in the ribosome and the spliceosome, and we will see others in Chapter 7 when we discuss *riboswitches*. One of the most dramatic RNA conformational changes has been observed with an artificial ribozyme which can exist in two entirely different conformations, each with a different catalytic activity (**Figure 6–105**). Since the discovery of catalysis by RNA, it has become clear that RNA is an enormously versatile molecule, and it is therefore not unreasonable to contemplate the past existence of an RNA world with a very high level of biochemical sophistication.

Self-Replicating Molecules Undergo Natural Selection

The three-dimensional folded structure of a polynucleotide affects its stability, its actions on other molecules, and its ability to replicate. Therefore, certain polynucleotides will be especially successful in any self-replicating mixture. Because errors inevitably occur in any copying process, new variant sequences of these polynucleotides will be generated over time.

Certain catalytic activities would have had a cardinal importance in the early evolution of life. Consider in particular an RNA molecule that helps to catalyze the process of templated polymerization, taking any given RNA molecule as a template (**Figure 6–106**). Such a molecule, by acting on copies of itself, can replicate. At the same time, it can promote the replication of other types of RNA molecules in its neighborhood (**Figure 6–107**). If some of these neighboring RNAs have catalytic actions that help the survival of RNA in other ways (catalyzing ribonucleotide production, for example), a set of different types of RNA molecules, each specialized for a different activity, could evolve into a cooperative system that replicates with unusually great efficiency.

But for any of these cooperative systems to evolve, they must be present together in a compartment. For example, a set of mutually beneficial RNAs (such

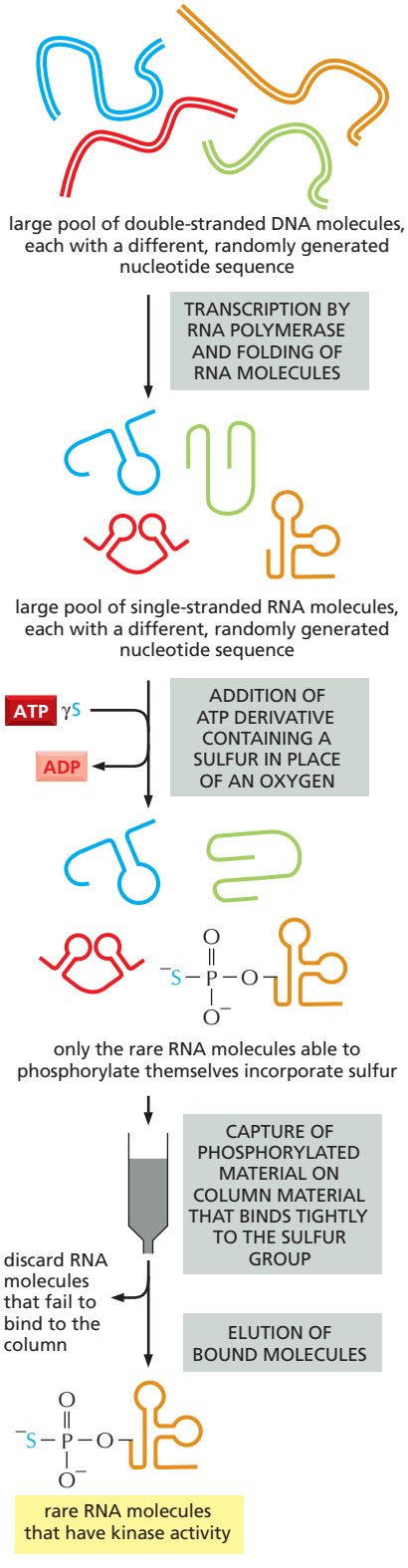


Figure 6–104 ***In vitro*** **selection of a synthetic ribozyme.** Beginning with a large pool of nucleic acid molecules synthesized in the laboratory, those rare RNA molecules that possess a specified catalytic activity can be isolated and studied. Although a specific example (that of an autophosphorylating ribozyme) is shown, variations of this procedure have been used to generate many of the ribozymes listed in Table 6–5. During the autophosphorylation step, the RNA molecules are kept sufficiently dilute to prevent the "cross"-phosphorylation of additional RNA molecules. In reality, several repetitions of this procedure are necessary to select the very rare RNA molecules with this catalytic activity. Thus, the material initially eluted from the column is converted back into DNA, amplified many fold (using reverse transcriptase and PCR as explained in Chapter 8), transcribed back into RNA, and subjected to repeated rounds of selection. (Adapted from J.R. Lorsch and J.W. Szostak, *Nature* 371:31–36, 1994. With permission from Macmillan Publishers Ltd.)

as those of Figure 6–107) could replicate themselves only if all the RNAs remained in the neighborhood of the RNA that is specialized for templated polymerization. Moreover, compartmentalization would bar parasitic RNA molecules from entering the system. Selection of a set of RNA molecules according to the quality of the self-replicating systems they generated could not therefore occur efficiently until some form of compartment evolved to contain them.

An early, crude form of compartmentalization may have been simple adsorption on surfaces or particles. The need for more sophisticated types of containment is easily fulfilled by a class of small molecules that has the simple physicochemical property of being *amphiphilic*, that is, consisting of one part that is hydrophobic (water insoluble) and another part that is hydrophilic (water soluble). When such molecules are placed in water, they aggregate, arranging their hydrophobic portions as much in contact with one another as possible and their hydrophilic portions in contact with the water. Amphiphilic molecules of appropriate shape aggregate spontaneously to form *bilayers*, creating small closed vesicles whose aqueous contents are isolated from the external medium (**Figure 6–108**). The phenomenon can be demonstrated in a test tube by simply mixing phospholipids and water together: under appropriate conditions, small vesicles will form. All present-day cells are surrounded by a *plasma membrane* consisting of amphiphilic molecules—mainly phospholipids—in this configuration; we discuss these molecules in detail in Chapter 10.

The spontaneous assembly of a set of amphiphilic molecules, enclosing a self-replicating mixture of RNAs (or pre-RNAs) and other molecules (**Figure**

Table 6–5 Some Biochemical Reactions That Can Be Catalyzed by Ribozymes

ACTIVITY	RIBOZYMES
Peptide bond formation in protein synthesis	ribosomal RNA
RNA cleavage, RNA ligation	self-splicing RNAs; RNase P; also *in vitro* selected RNA
DNA cleavage	self-splicing RNAs
RNA splicing	self-splicing RNAs, perhaps RNAs of the spliceosome
RNA polymerizaton	*in vitro* selected RNA
RNA and DNA phosphorylation	*in vitro* selected RNA
RNA aminoacylation	*in vitro* selected RNA
RNA alkylation	*in vitro* selected RNA
Amide bond formation	*in vitro* selected RNA
Glycosidic bond formation	*in vitro* selected RNA
Oxidation/reduction reactions	*in vitro* selected RNA
Carbon–carbon bond formation	*in vitro* selected RNA
Phosphoamide bond formation	*in vitro* selected RNA
Disulfide exchange	*in vitro* selected RNA

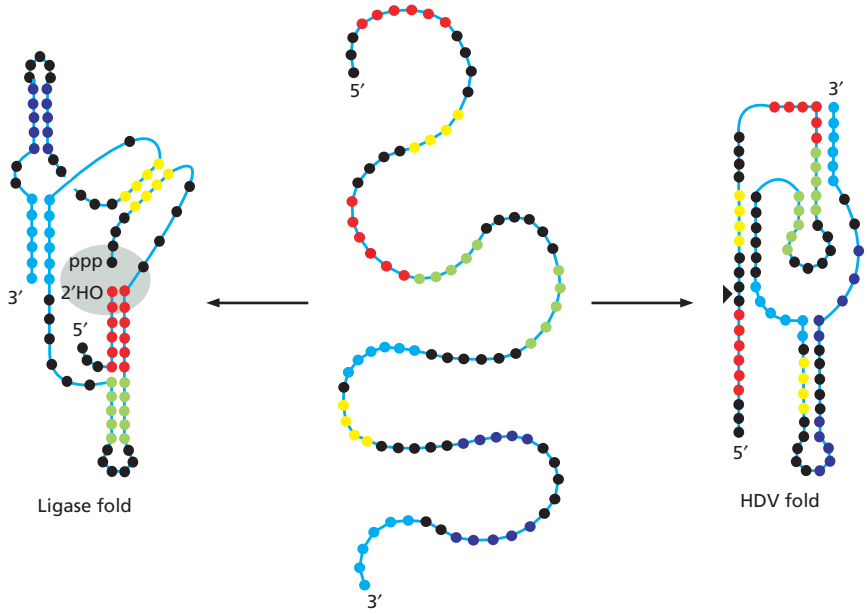


Figure 6–105 An RNA molecule that folds into two different ribozymes. This 88-nucleotide RNA, created in the laboratory, can fold into a ribozyme that carries out a self-ligation reaction *(left)* or a self-cleavage reaction *(right)*. The ligation reaction forms a 2′,5′ phosphodiester linkage with the release of pyrophosphate. This reaction seals the gap *(gray* shading), which was experimentally introduced into the RNA molecule. In the reaction carried out by the HDV fold, the RNA is cleaved at this same position, indicated by the *arrowhead*. This cleavage resembles that used in the life cycle of HDV, a hepatitis B satellite virus, hence the name of the fold. Each nucleotide is represented by a *colored dot*, with the colors used simply to clarify the two different folding patterns. The folded structures illustrate the secondary structures of the two ribozymes with regions of base-pairing indicated by close oppositions of the *colored dots*. Note that the two ribozyme folds have no secondary structure in common. (Adapted from E.A. Schultes and D.P. Bartel, *Science* 289:448–452, 2000. With permission from AAAS.)

6–109), presumably formed the first membrane-bounded cells. Although it is not clear at what point in the evolution of biological catalysts this might have occurred, once RNA molecules were sealed within a closed membrane they could begin to evolve in earnest as carriers of genetic instructions: new variants could be selected not merely on the basis of their own structure, but also according to their effect on the other molecules in the same compartment. The nucleotide sequences of the RNA molecules could now be expressed in the character of a unitary living cell.

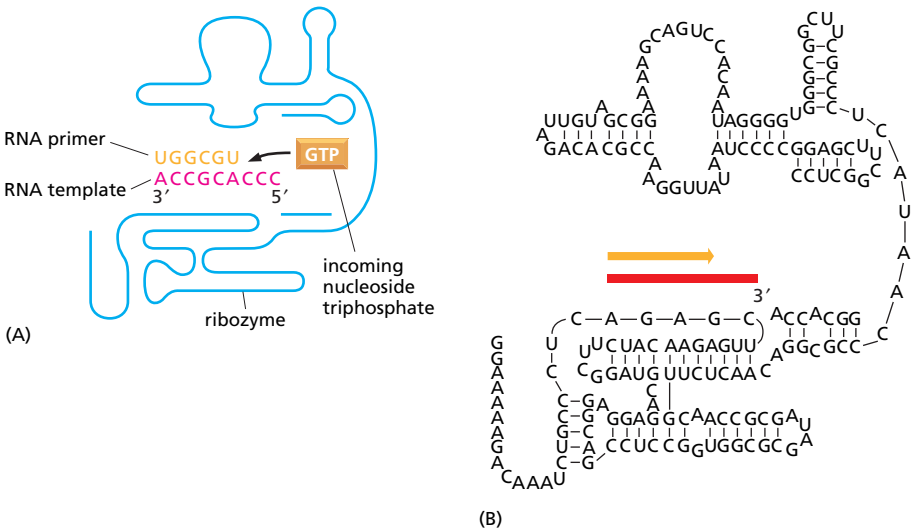


Figure 6–106 A ribozyme created in the laboratory that can catalyze templated synthesis of RNA from nucleoside triphosphates. (A) Schematic diagram of the ribozyme showing one step of the templated polymerization reaction it catalyzes. (B) Nucleotide sequence of the ribozyme with base pairings indicated. Although relatively inefficient (it can only synthesize short lengths of RNA), this ribozyme adds the correct base, as specified by the template, over 95% of the time. (From W.K. Johnston et al., *Science* 292:1319–1325, 2001. With permission from AAAS.)

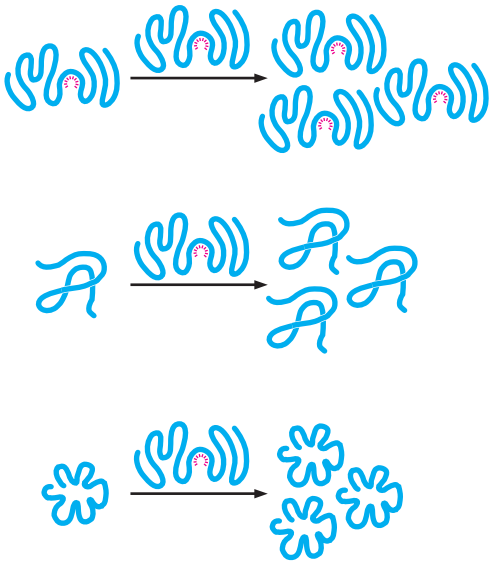


Figure 6–107 A family of mutually supportive RNA molecules. One molecule is a ribozyme that replicates itself as well as the other RNA molecules. The other molecules would catalyze secondary tasks needed for the survival of the cooperative system, for example, by synthesizing ribonucleotides for RNA synthesis or phospholipids for compartmentalization.

How Did Protein Synthesis Evolve?

The molecular processes underlying protein synthesis in present-day cells seem inextricably complex. Although we understand most of them, they do not make conceptual sense in the way that DNA transcription, DNA repair, and DNA replication do. It is especially difficult to imagine how protein synthesis evolved because it is now performed by a complex interlocking system of protein and RNA molecules; obviously the proteins could not have existed until an early version of the translation apparatus was already in place. The RNA world hypothesis is especially appealing because the use of RNA in both information and catalysis seems both economic and conceptually simple. As attractive as this idea is for envisioning early life, it does not explain how the modern-day system of protein synthesis arose. Although we can only speculate on the origins of modern protein synthesis and the genetic code, several experimental observations have provided plausible scenarios.

In modern cells, some short peptides (such as antibiotics) are synthesized without the ribosome; peptide synthetase enzymes assemble these peptides, with their proper sequence of amino acids, without mRNAs to guide their synthesis. It is plausible that this non-coded, primitive version of protein synthesis first developed during the RNA world where it would have been catalyzed by RNA molecules. This idea presents no conceptual difficulties because, as we have seen, rRNA catalyzes peptide bond formation in present-day cells. We also know that ribozymes created in the laboratory can perform specific aminoacylation reactions; that is, they can match specific amino acids to specific tRNAs. It is therefore possible that tRNA-like adapters, each matched to a specific amino acid, could have arisen in the RNA world, marking the beginnings of a genetic code.

In principle, other RNAs (the precursors to mRNAs) could have served as crude templates to direct the nonrandom polymerization of a few different amino acids. Any RNA that helped guide the synthesis of a useful polypeptide would have a great advantage in the evolutionary struggle for survival. We can envision a relatively nonspecific peptidyl transferase ribozyme, which, over time, grew larger and acquired the ability to position charged tRNAs accurately on RNA templates—leading eventually to the modern ribosome. Once protein

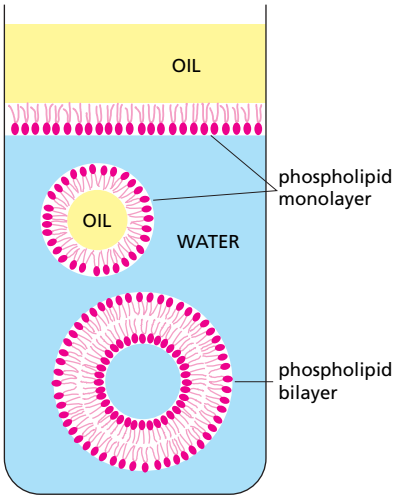


Figure 6–108 Formation of membrane by phospholipids. Because these molecules have hydrophilic heads and lipophilic tails, they align themselves at an oil/water interface with their heads in the water and their tails in the oil. In the water they associate to form closed bilayer vesicles in which the lipophilic tails are in contact with one another and the hydrophilic heads are exposed to the water.

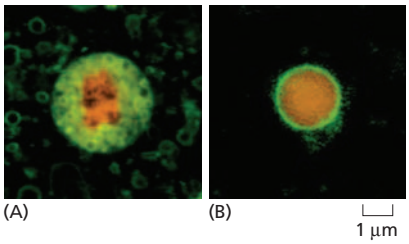


Figure 6–109 Encapsulation of RNA by simple amphiphilic molecules. For these experiments, the clay mineral montmorillonite was used to bring together RNA and fatty acids. (A) A montmorillonite particle, coated by RNA (*red*) has become trapped inside a fatty acid vesicle (*green*). (B) RNA *(red)* in solution has been encapsulated by fatty acids *(green)*. These experiments show that montmorillonite can greatly accelerate the spontaneous generation of vesicles from amphiphilic molecules and trap RNA inside them. It has been hypothesized that conceptually similar actions may have led to the first primitive cells on Earth. (From M.M. Hanczyc et al., *Science* 302:618–622, 2003. With permission from AAAS.)

synthesis evolved, the transition to a protein-dominated world could proceed, with proteins eventually taking over the majority of catalytic and structural tasks because of their greater versatility, with 20 rather than 4 different subunits. Although the scenarios just discussed are highly speculative, the known properties of RNA molecules are consistent with these ideas.

All Present-Day Cells Use DNA as Their Hereditary Material

If the evolutionary speculations embodied in the RNA world hypothesis are correct, early cells would have differed fundamentally from the cells we know today in having their hereditary information stored in RNA rather than in DNA (**Figure 6–110**). Evidence that RNA arose before DNA in evolution can be found in the chemical differences between them. Ribose, like glucose and other simple carbohydrates, can be formed from formaldehyde (HCHO), a simple chemical which is readily produced in laboratory experiments that attempt to simulate conditions on the primitive Earth. The sugar deoxyribose is harder to make, and in present-day cells it is produced from ribose in a reaction catalyzed by a protein enzyme, suggesting that ribose predates deoxyribose in cells. Presumably, DNA appeared on the scene later, but then proved more suitable than RNA as a permanent repository of genetic information. In particular, the deoxyribose in its sugar-phosphate backbone makes chains of DNA chemically more stable than chains of RNA, so that much greater lengths of DNA can be maintained without breakage.

The other differences between RNA and DNA—the double-helical structure of DNA and the use of thymine rather than uracil—further enhance DNA stability by making the many unavoidable accidents that occur to the molecule much easier to repair, as discussed in detail in Chapter 5 (see pp. 296–297 and 300–301).

Summary

From our knowledge of present-day organisms and the molecules they contain, it seems likely that the development of the directly autocatalytic mechanisms fundamental to living systems began with the evolution of families of molecules that could catalyze their own replication. With time, a family of cooperating RNA catalysts probably developed the ability to direct the synthesis of polypeptides. DNA is likely to have been a late addition: as the accumulation of additional protein catalysts allowed more efficient and complex cells to evolve, the DNA double helix replaced RNA as a more stable molecule for storing the increased amounts of genetic information required by such cells.

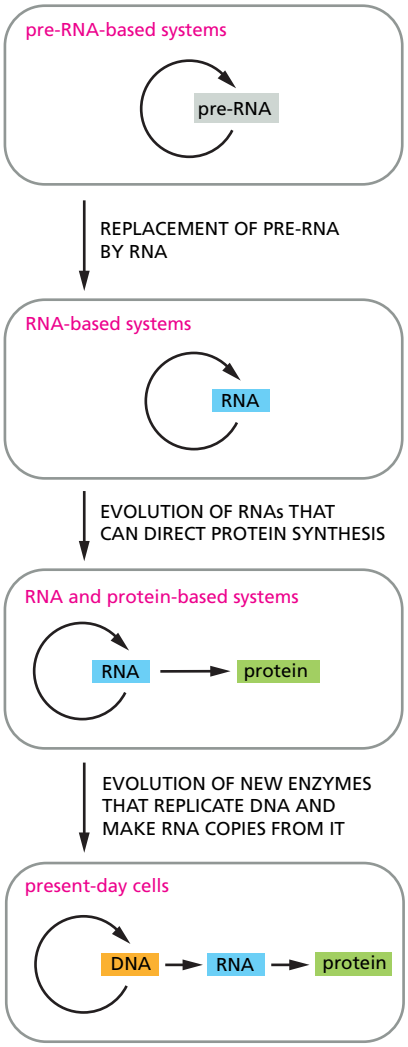


Figure 6–110 The hypothesis that RNA preceded DNA and proteins in evolution. In the earliest cells, pre-RNA molecules would have had combined genetic, structural, and catalytic functions and RNA would have gradually taken over these functions. In present-day cells, DNA is the repository of genetic information, and proteins perform the vast majority of catalytic functions in cells. RNA primarily functions today as a go-between in protein synthesis, although it remains a catalyst for a small number of crucial reactions.

PROBLEMS

Which statements are true? Explain why or why not.

6–1 The consequences of errors in transcription are less than those of errors in RNA replication.

6–2 Since introns are largely genetic "junk," they do not have to be removed precisely from the primary transcript during RNA splicing.

6–3 Wobble pairing occurs between the first position in the codon and the third position in the anticodon.

6–4 Protein enzymes are thought to greatly outnumber ribozymes in modern cells because they catalyze a much greater variety of reactions at much faster rates than ribozymes.

Discuss the following problems.

6–5 In which direction along the template must the RNA polymerase in **Figure Q6–1** be moving to have generated the supercoiled structures that are shown? Would you expect supercoils to be generated if the RNA polymerase were free to rotate about the axis of the DNA as it progressed along the template?

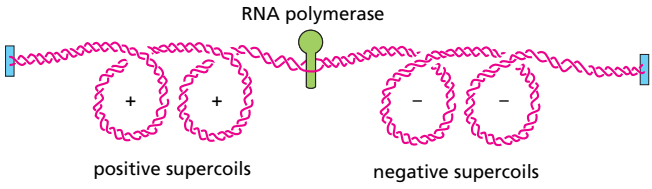


Figure Q6–1 Supercoils around a moving RNA polymerase (Problem 6–5).

6–6 Phosphates are attached to the CTD (C-terminal domain) of RNA polymerasse II during transcription. What are the various roles of RNA polymerase II CTD phosphorylation?

6–7 The human α-tropomyosin gene is alternatively spliced to produce several forms of α-tropomyosin mRNA in various cell types (**Figure Q6–2**). For all forms of the mRNA, the encoded protein sequence is the same for exons 1 and

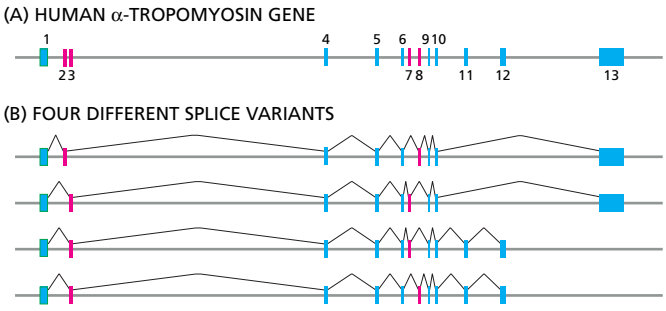


Figure Q6–2 Alternatively spliced mRNAs from the human α-tropomyosin gene (Problem 6–7). (A) Exons in the human α-tropomyosin gene. The locations and relative sizes of exons are shown by the *blue* and *red rectangles*. (B) Splicing patterns for four α-tropomyosin mRNAs. Splicing is indicated by *lines* connecting the exons that are included in the mRNA.

10. Exons 2 and 3 are alternative exons used in different mRNAs, as are exons 7 and 8. Which of the following statements about exons 2 and 3 is the most accurate? Is that statement also the most accurate one for exons 7 and 8? Explain your answers.

A. Exons 2 and 3 must have the same number of nucleotides.

B. Exons 2 and 3 must each contain an integral number of codons (that is, the number of nucleotides divided by 3 must be an integer).

C. Exons 2 and 3 must each contain a number of nucleotides that when divided by 3 leaves the same remainder (that is, 0, 1, or 2).

6–8 After treating cells with a chemical mutagen, you isolate two mutants. One carries alanine and the other carries methionine at a site in the protein that normally contains valine (**Figure Q6–3**). After treating these two mutants again with the mutagen, you isolate mutants from each that now carry threonine at the site of the original valine (Figure Q6–3). Assuming that all mutations involve single nucleotide changes, deduce the codons that are used for valine, methionine, threonine, and alanine at the affected site. Would you expect to be able to isolate valine-to-threonine mutants in one step?

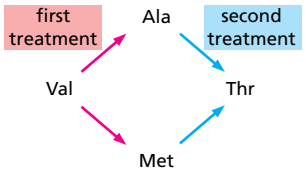


Figure Q6–3 Two rounds of mutagenesis and the altered amino acids at a single position in a protein (Problem 6–8).

6–9 The elongation factor EF-Tu introduces two short delays between codon–anticodon base-pairing and formation of the peptide bond. These delays increase the accuracy of protein synthesis. Describe these delays and explain how they improve the fidelity of translation.

6–10 Both Hsp60-like and Hsp70 molecular chaperones share an affinity for exposed hydrophobic patches on proteins, using them as indicators of incomplete folding. Why do you suppose hydrophobic patches serve as critical signals for the folding status of a protein?

6–11 Most proteins require molecular chaperones to assist in their correct folding. How do you suppose the chaperones themselves manage to fold correctly?

6–12 What is so special about RNA that makes it such an attractive evolutionary precursor to DNA and protein? What is it about DNA that makes it a better material than RNA for storage of genetic information?

6–13 If an RNA molecule could form a hairpin with a symmetric internal loop, as shown in **Figure Q6–4**, could the complement of this RNA form a similar structure? If so, would there be any regions of the two structures that are identical? Which ones?

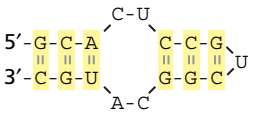


Figure Q6–4 An RNA hairpin with a symmetric internal loop (Problem 6–13).

REFERENCES

General

Berg JM, Tymoczko JL & Stryer L (2006) Biochemistry, 6th ed. New York: WH Freeman.
Brown TA (2002) Genomes 2, 2nd ed. New York: Wiley-Liss.
Gesteland RF, Cech TR & Atkins JF (eds) (2006) The RNA World, 3rd ed. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.
Hartwell L, Hood L, Goldberg ML et al (2006) Genetics: from Genes to Genomes, 3rd ed. Boston: McGraw Hill.
Lodish H, Berk A, Kaiser C et al (2007) Molecular Cell Biology, 6th ed. New York: WH Freeman.
Stent GS (1971) Molecular Genetics: An Introductory Narrative. San Francisco: WH Freeman.
The Genetic Code (1966) *Cold Spring Harb. Symp Quant Biol* 31.
The Ribosome (2001) *Cold Spring Harb Symp Quant Biol* 66.
Watson JD, Baker TA, Bell SP et al (2003) Molecular Biology of the Gene, 5th ed. Menlo Park, CA: Benjamin Cummings.

From DNA to RNA

Bentley DL (2005) Rules of engagement: co-transcriptional recruitment of pre-mRNA processing factors. *Curr Opin Cell Biol* 17:251–256.
Berget SM, Moore C & Sharp PA (1977) Spliced segments at the 5′ terminus of adenovirus 2 late mRNA. *Proc Natl Acad Sci USA* 74:3171–3175.
Black DL (2003). Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem* 72:291–336.
Brenner S, Jacob F & Meselson M (1961) An unstable intermediate carrying information from genes to ribosomes for protein synthesis. *Nature* 190:576–581.
Cate JH, Gooding AR, Podell E et al. (1996) Crystal structure of a group I ribozyme domain: principles of RNA packing. *Science* 273:1678–1685.
Chow LT, Gelinas RE, Broker TR et al (1977) An amazing sequence arrangement at the 5′ ends of adenovirus 2 messenger RNA. *Cell* 12:1–8.
Cramer P (2002) Multisubunit RNA polymerases. *Curr Opin Struct Biol* 12:89–97.
Daneholt B (1997) A look at messenger RNP moving through the nuclear pore. *Cell* 88:585–588.
Dreyfuss G, Kim VN, & Kataoka N (2002) Messenger-RNA-binding proteins and the messages they carry. *Nature Rev Mol Cell Biol* 3:195–205.
Houseley J, LaCava J & Tollervey D (2006) RNA-quality control by the exosome. *Nature Rev Mol Cell Biol* 7:529–539.
Izquierdo JM, & Valcárcel J (2006) A simple principle to explain the evolution of pre-mRNA splicing. *Genes Dev* 20:1679–1684.
Kornberg RD (2005) Mediator and the mechanism of transcriptional activation. *Trends Biochem Sci* 30:235–239.
Malik S & Roeder RG (2005) Dynamic regulation of pol II transcription by the mammalian Mediator complex. *Trends Biochem Sci* 30:256–263.
Matsui T, Segall J, Weil PA & Roeder RG (1980) Multiple factors required for accurate initiation of transcription by purified RNA polymerase II. *J Biol Chem* 255:11992–11996.
Patel AA & Steitz JA (2003) Splicing double: insights from the second spliceosome. *Nature Rev Mol Cell Biol* 4|:960–970.
Phatnani HP & Greenleaf AL (2006) Phosphorylation and functions of the RNA polymerase II CTD. *Genes Dev* 20:2922–2936.
Query CC & Konarska MM (2006) Splicing fidelity revisited. *Nature Struct Mol Biol* 13:472–474.
Ruskin B, Krainer AR, Maniatis T et al (1984) Excision of an intact intron as a novel lariat structure during pre-mRNA splicing *in vitro*. *Cell* 38:317–331.
Spector DL (2003) The dynamics of chromosome organization and gene regulation. *Annu Rev Biochem* 72:573–608.
Staley JP & Guthrie C (1998) Mechanical devices of the spliceosome: motors, clocks, springs, and things. *Cell* 92:315–326.
Thomas MC & Chiang CM (2006) The general transcription machinery and general cofactors. *Critical Rev Biochem Mol Biol* 41:105–178.
Wang D, Bushnell DA, Westover KD et al (2006) Structural basis of transcription: role of the trigger loop in substrate specificity and catalysis. *Cell* 127:941–954.

From RNA to Protein

Allen GS & Frank J (2007) Structural insights on the translation initiation complex: ghosts of a universal initiation complex. *Mol Microbiol* 63:941–950.
Anfinsen CB (1973) Principles that govern the folding of protein chains. *Science* 181:223–230.
Brunelle JL, Youngman EM, Sharma D et al (2006) The interaction between C75 of tRNA and the A loop of the ribosome stimulates peptidyl transferase activity. *RNA* 12:33–39.
Chien P, Weissman JS, & DePace AH (2004). Emerging principles of conformation-based prion inheritance. *Annu Rev Biochem* 73:617–656.
Crick FHC (1966) The genetic code: III. *Sci Am* 215:55–62.
Hershko A, Ciechanover A & Varshavsky A (2000) The ubiquitin system. *Nature Med* 6:1073–1081.
Ibba M & Soll D (2000) Aminoacyl-tRNA synthesis. *Annu Rev Biochem* 69:617–650.
Kozak M (1992) Regulation of translation in eukaryotic systems. *Annu Rev Cell Biol* 8:197–225.
Kuzmiak HA, & Maquat LE (2006) Applying nonsense-mediated mRNA decay research to the clinic: progress and challenges. *Trends Mol Med* 12:306–316.
Moore PB & Steitz TA (2005) The ribosome revealed. *Trends Biochem Sci* 30:281–283.
Noller HF (2005) RNA structure: reading the ribosome. *Science* 309:1508–1514.
Ogle JM, Carter AP & Ramakrishnan V (2003) Insights into the decoding mechanism from recent ribosome structures. *Trends Biochem Sci* 28:259–266.
Prusiner SB (1998) Nobel lecture. Prions. *Proc Natl Acad Sci USA* 95:13363–13383.
Rehwinkel J, Raes J & Izaurralde E (2006) Nonsense-mediated mRNA decay: Target genes and functional diversification of effectors. *Trends Biochem Sci* 31:639–646.
Sauer RT, Bolon DN, Burton BM et al (2004) Sculpting the proteome with AAA(+) proteases and disassembly machines. *Cell* 119:9–18.
Shorter J & Lindquist S (2005) Prions as adaptive conduits of memory and inheritance. *Nature Rev Genet* 6:435–450.
Varshavsky A (2005) Regulated protein degradation. *Trends in Biochem Sci* 30:283–286.
Voges D, Zwickl P & Baumeister W (1999) The 26S proteasome: a molecular machine designed for controlled proteolysis. *Annu Rev Biochem* 68:1015–1068.
Weissmann C (2005) Birth of a prion: spontaneous generation revisited. *Cell* 122:165–168.
Young JC, Agashe VR, Siegers K et al (2004) Pathways of chaperone-mediated protein folding in the cytosol. *Nature Rev Mol Cell Biol* 5:781–791.

The RNA World and the Origins of Life

Joyce GF (1992) Directed molecular evolution. *Sci Am* 267:90–97.
Orgel L (2000) Origin of life. A simpler nucleic acid. *Science* 290:1306–1307.
Kruger K, Grabowski P, Zaug P et al (1982) Self-splicing RNA: Autoexcision and autocyclization of the ribosomal RNA intervening seuence of Tetrahymena. *Cell* 31:147–157.
Silverman SK (2003) Rube Goldberg goes (ribo)nuclear? Molecular switches and sensors made from RNA. *RNA* 9:377–383.
Szostak JW, Bartel DP & Luisi PL (2001) Synthesizing life. *Nature* 409:387–390.