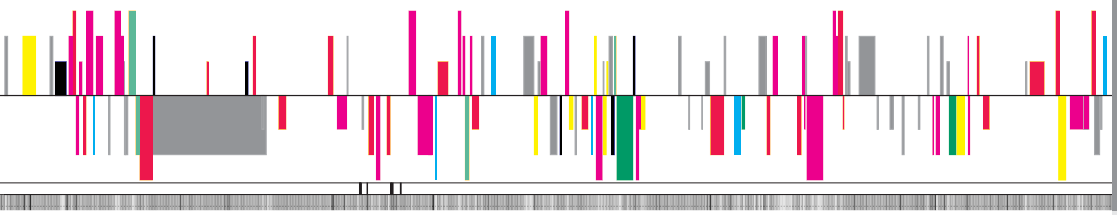


6

HOW CELLS READ THE GENOME: FROM DNA TO PROTEIN

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. Fifty years later, we have complete genome sequences for many organisms, including humans, and 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. We shall see that 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 fruitfly, or a human. Nevertheless, we still have a great deal to discover about how the information stored in an organism's genome produces even the simplest unicellular bacterium with 500 genes, let alone how it directs the development of a human with approximately 30,000 genes. An enormous amount of ignorance remains; many fascinating challenges therefore await the next generation of cell biologists.

The problems 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 is used to specify 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 particular protein is made by the cell, the corresponding region of the genome must therefore be accurately decoded. 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.

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: *rainbow-colored bar*: G–C base-pair content; *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 predicted gene includes both its exons (protein-coding DNA) and its introns (non-coding DNA) (see Figure 4–25). As indicated in the key, the height of each gene box is proportional to the number of cDNAs in various databases 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 higher the number of matches between the nucleotide sequences of cDNAs and that of a particular predicted gene, the higher the confidence that the predicted gene is transcribed into RNA and is thus a genuine gene. The color of each gene box (see *color code* in the key) 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. (From Mark D. Adams et al., *Science* 287:2185–2195, 2000. © AAAS.)

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 for humans 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.

The DNA in genomes does not direct protein synthesis itself, but instead uses RNA as an intermediary molecule. 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 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 the genome. Finally, although we focus on the production of the proteins encoded by the genome in this chapter, we see that for some genes RNA is the final product. Like proteins, many of these RNAs fold into precise three-dimensional structures that have structural and catalytic 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 cellular evolution.

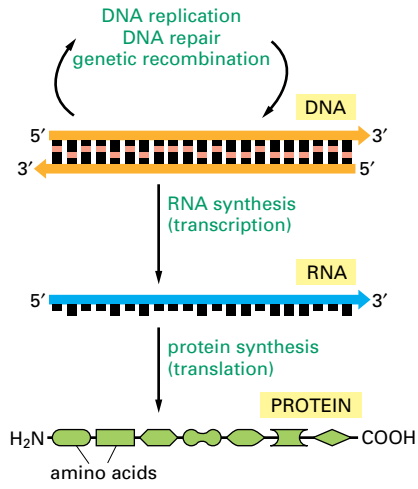


Figure 6–2 The pathway from DNA to protein. The flow of genetic information from DNA to RNA (transcription) and from RNA to protein (translation) occurs in all living cells.

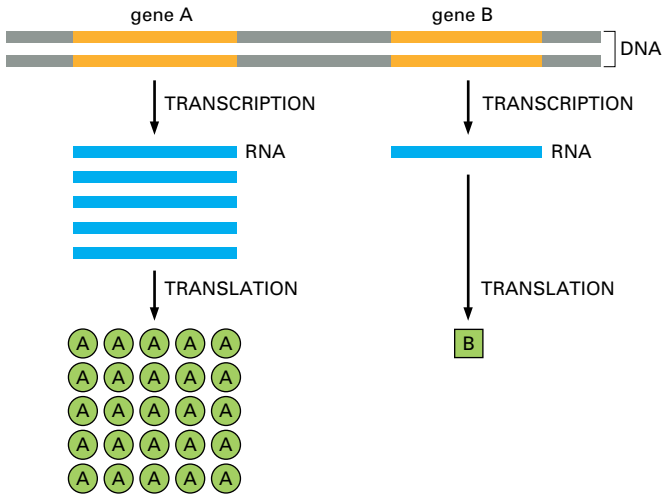


Figure 6–3 Genes can be expressed with different efficiencies. Gene A is transcribed and translated much more efficiently than gene B. This allows the amount of protein A in the cell to be much greater than that of protein B.

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 obviously by controlling the production of its RNA.

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). It is not uncommon, however, to find other types of base pairs in RNA: for example, G pairing with U occasionally.

Despite these small chemical differences, DNA and RNA differ quite dramatically in overall structure. Whereas DNA always occurs in cells as a double-stranded helix, RNA is single-stranded. RNA chains therefore fold up into a variety of shapes, 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 structural and catalytic functions.

Transcription Produces RNA Complementary to One Strand of DNA

All of 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.

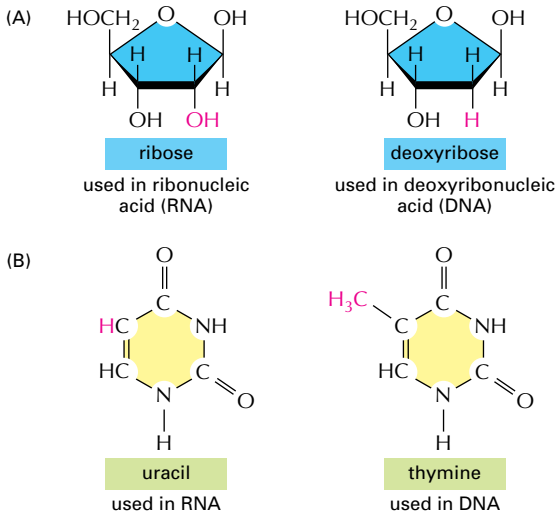
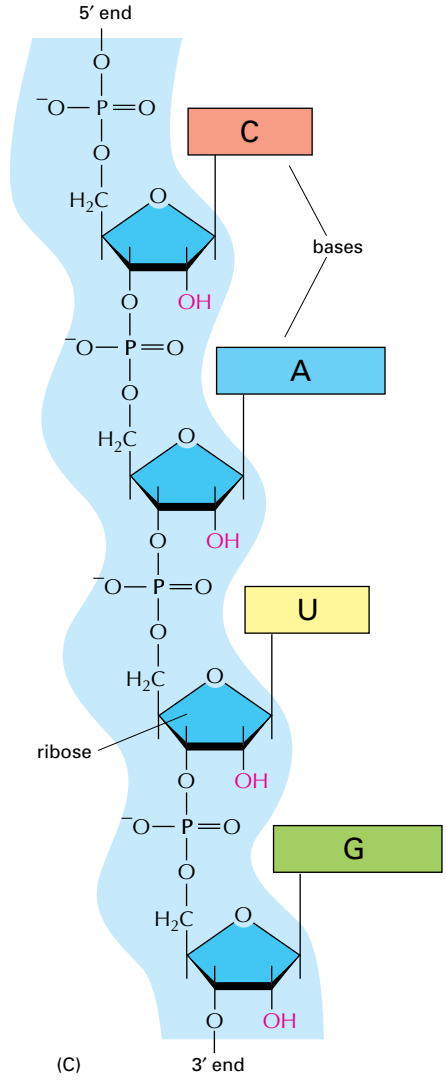


Figure 6–4 The chemical structure of RNA. (A) RNA contains the sugar ribose, which differs from deoxyribose, the sugar used in DNA, by the presence of an additional –OH group. (B) RNA contains the base uracil, which differs from thymine, the equivalent base in DNA, by the absence of a $–CH_3$ group.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



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.

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 for DNA replication, a hydrolysis of high-energy bonds provides the energy needed to drive the reaction forward (see Figure 5–4).

The almost immediate release of the RNA strand from the DNA as it is synthesized means that many RNA copies can be made from the same gene in a

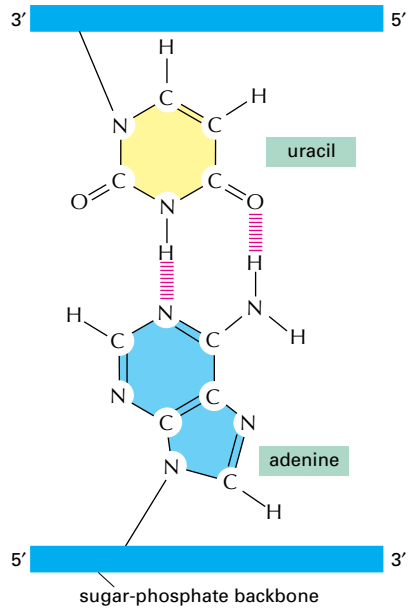


Figure 6–5 Uracil forms base pairs with adenine. The absence of a methyl group in U has no effect on base-pairing; thus, U–A base pairs closely resemble T–A base pairs (see Figure 4–4).

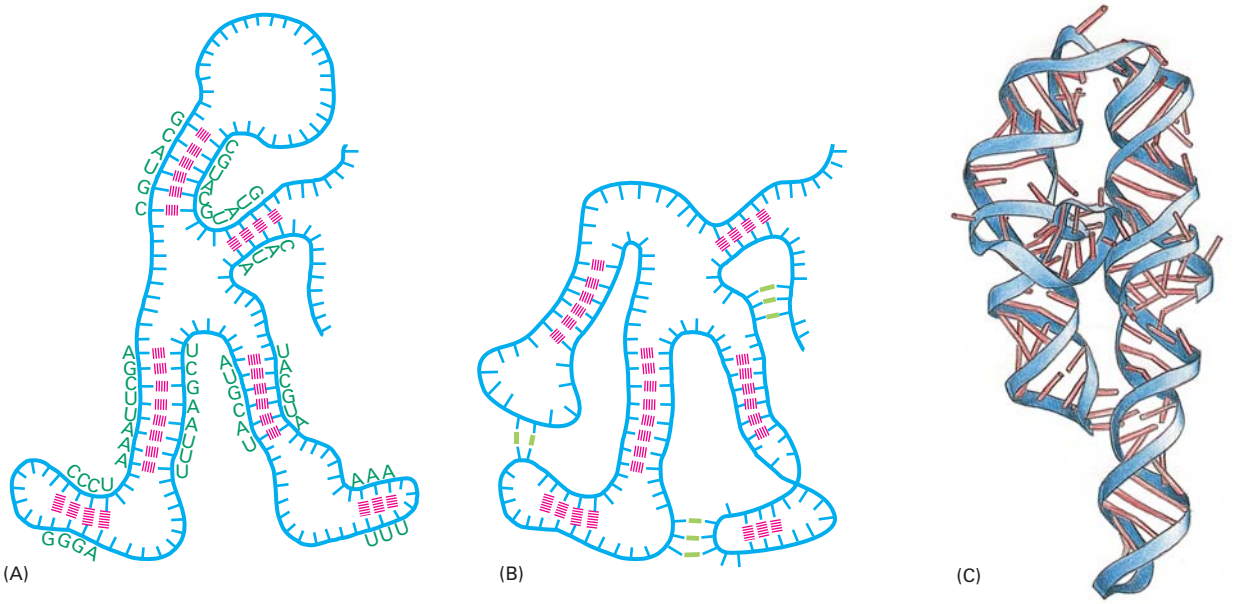


Figure 6–6 RNA can fold into specific structures. RNA is largely single-stranded, but it often contains short stretches of nucleotides that can form conventional base-pairs with complementary sequences found elsewhere on the same molecule. These interactions, along with additional "nonconventional" base-pair interactions, allow an RNA molecule to fold into a three-dimensional structure that is determined by its sequence of nucleotides. (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.

relatively short time, 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 two enzymes. First, and most obvious, 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. 243). 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 the 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 mimics the reverse of the polymerization reaction, except that water instead of pyrophosphate is used (see Figure 5–4). RNA polymerase hovers around a misincorporated ribonucleotide longer than it does for a correct addition, causing excision to be favored for incorrect nucleotides. However, RNA polymerase also excises many correct bases as part of the cost for improved accuracy.

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.

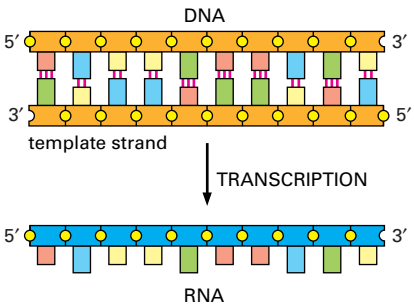


Figure 6–7 DNA transcription produces a single-stranded RNA molecule that is complementary to one strand of DNA.

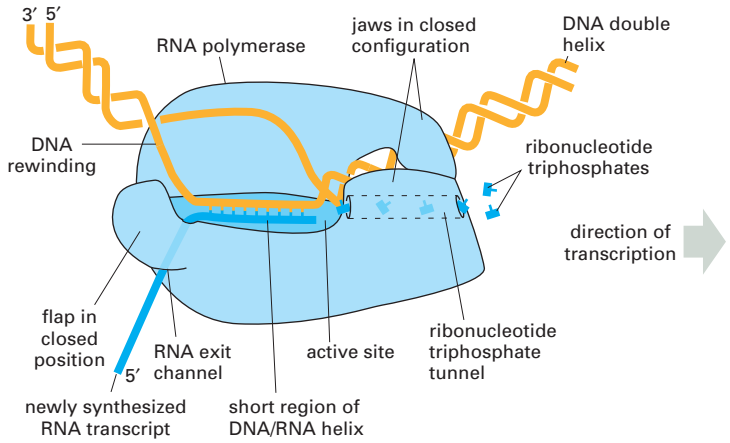


Figure 6–8 DNA is transcribed by the enzyme RNA polymerase. The RNA polymerase *(pale blue)* moves stepwise along the DNA, unwinding the DNA helix at its active site. As it progresses, the polymerase adds nucleotides (here, *small "T" shapes)* one by one to the RNA chain at the polymerization site using an exposed DNA strand as a template. The RNA transcript is thus a single-stranded complementary copy of one of the two DNA strands. The polymerase has a rudder (see Figure 6–11) that displaces the newly formed RNA, allowing the two strands of DNA behind the polymerase to rewind. A short region of DNA/RNA helix (approximately nine nucleotides 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). (Adapted from a figure kindly supplied by Robert Landick.)

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, although this number includes multiple copies of some highly repeated genes. These RNAs, like proteins, serve as enzymatic and structural components for a wide variety of processes in the cell. In Chapter 5 we encountered one of those RNAs, the template carried by the enzyme telomerase. Although not all of their functions are known, we see in this chapter that some *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 (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.

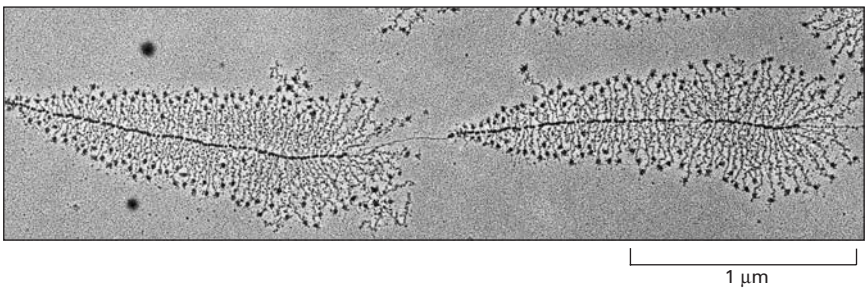


Figure 6–9 Transcription of two genes as observed under the electron microscope. The micrograph shows many molecules of RNA polymerase simultaneously transcribing each of two adjacent genes. Molecules of RNA polymerase are visible as a series of dots along the DNA with the newly synthesized transcripts (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.)

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
Other noncoding RNAs	function in diverse cellular processes, including telomere synthesis, X-chromosome inactivation, and the transport of proteins into the ER

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 process in bacteria is simpler, we look there 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. Bacterial RNA polymerase is a multisubunit complex. A detachable subunit, called *sigma (σ) factor*, is largely responsible for its ability to read the signals in the DNA that tell it where to begin transcribing (Figure 6–10). RNA polymerase molecules adhere only weakly to the bacterial DNA when they collide with it, and a polymerase molecule typically slides rapidly along the long DNA molecule until it dissociates again. However, when the polymerase slides into a region on the DNA double helix called a **promoter**, a special sequence of nucleotides indicating the starting point for RNA synthesis, it binds tightly to it. The polymerase, using its σ factor, recognizes this 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 6–10).

After the RNA polymerase 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 6–10). Unlike a DNA helicase reaction (see Figure 5–15), 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 more energetically favorable state. With the DNA unwound, one of the two exposed DNA strands acts as a template for complementary base-pairing with incoming ribonucleotides (see Figure 6–7), two of which are joined together by the polymerase to begin an RNA chain. After the first ten or so nucleotides of RNA have been synthesized (a relatively inefficient process during which polymerase synthesizes and discards short nucleotide oligomers), the σ factor relaxes its tight hold on the polymerase and evenutally dissociates from it. During this process, the polymerase undergoes additional structural changes that enable it to move forward rapidly, transcribing without the σ factor *(Step 4* in Figure 6–10). 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 DNA template and the newly made RNA chain *(Step 7* in Figure 6–10). After the polymerase has been released at a terminator, it reassociates with a free σ factor and searches for a new promoter, where it can begin the process of transcription again.

Several structural features of bacterial RNA polymerase make it particularly adept at performing the transcription cycle just described. Once the σ factor

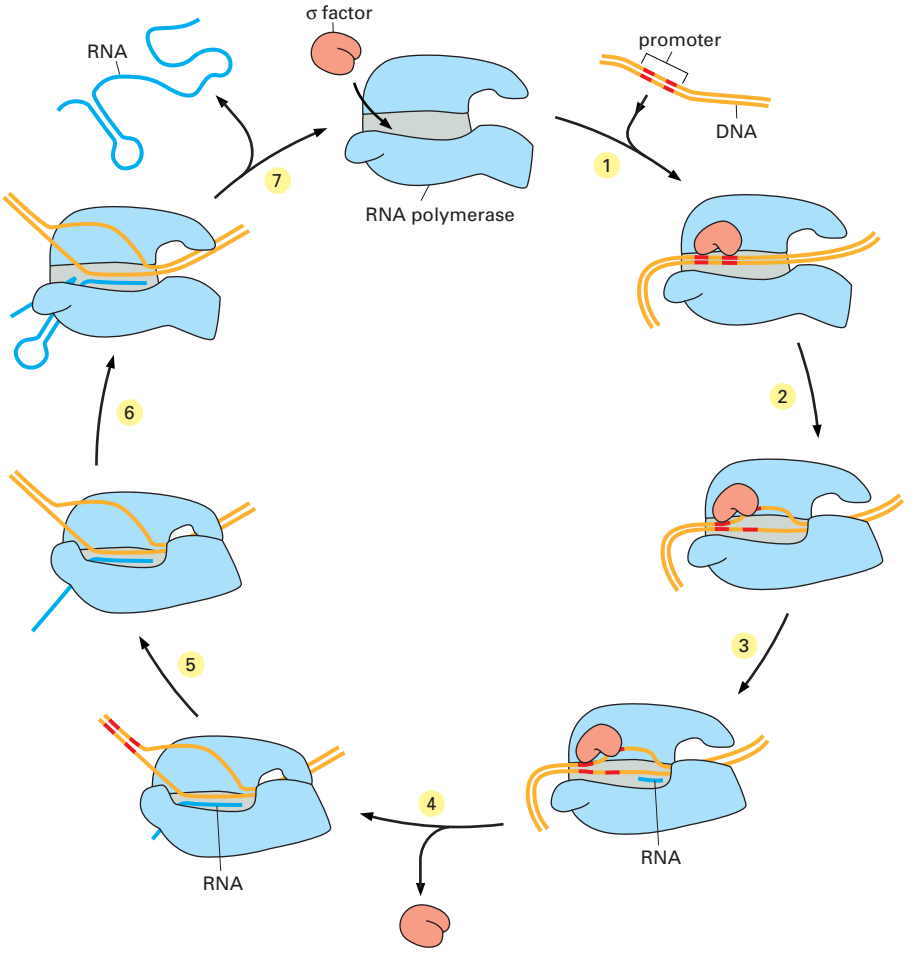


Figure 6–10 The transcription cycle of bacterial RNA polymerase. In step 1, the RNA polymerase holoenzyme (core polymerase plus σ factor) forms and then locates a promoter (see Figure 6–12). The polymerase unwinds the DNA at the position at which transcription is to begin (step 2) and begins transcribing (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, σ relaxes its grip, and the polymerase undergoes a series of conformational changes (which probably includes a tightening of its jaws and the placement of RNA in the exit channel [see Figure 6–11]). The polymerase now shifts to the elongation mode of RNA synthesis (step 4), moving rightwards 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 (step 6). Termination signals are encoded in DNA and many function by forming an RNA structure that destabilizes the polymerase's hold on the RNA, as shown here. 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 kindly supplied by Robert Landick.)

positions the polymerase on the promoter and the template DNA has been unwound and pushed to the active site, a pair of moveable jaws is thought to clamp onto the DNA (Figure 6–11). When the first 10 nucleotides have been transcribed, the dissociation of σ allows a flap at the back of the polymerase to

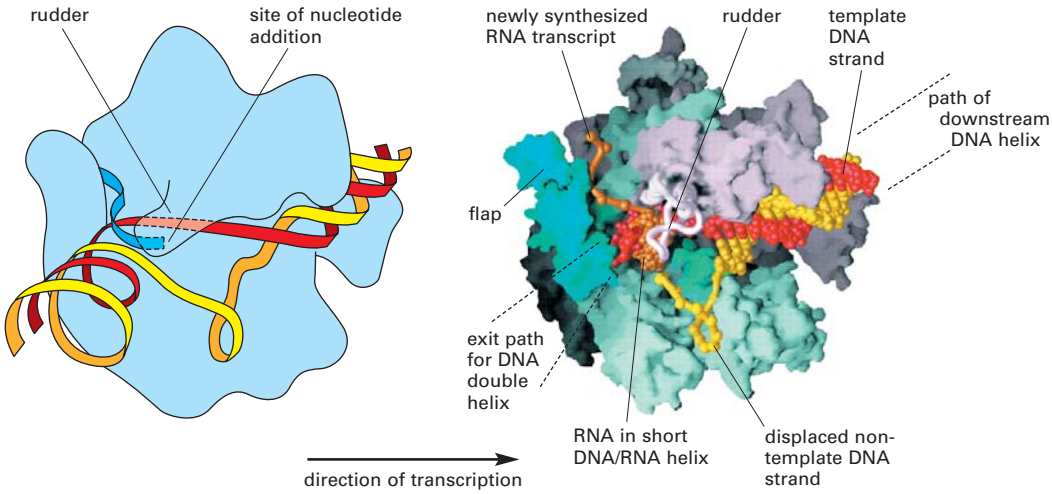
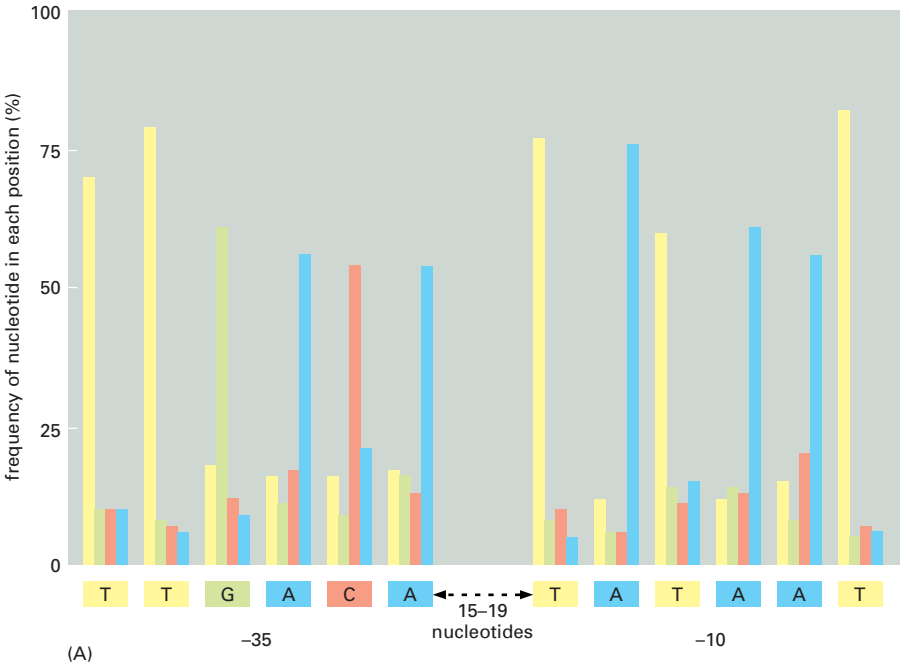


Figure 6–11 The structure of a bacterial RNA polymerase. Two depictions of the three-dimensional structure of a bacterial RNA polymerase, with the DNA and RNA modeled in. This RNA polymerase is formed from four different subunits, indicated by different colors *(right)*. The DNA strand used as a template is *red*, and the non-template strand is *yellow*. The rudder wedges apart the DNA–RNA hybrid as the polymerase moves. For simplicity only the polypeptide backbone of the rudder is shown in the right-hand figure, and the DNA exiting from the polymerase has been omitted. Because the RNA polymerase is depicted in the elongation mode, the σ factor is absent. (Courtesy of Seth Darst.)



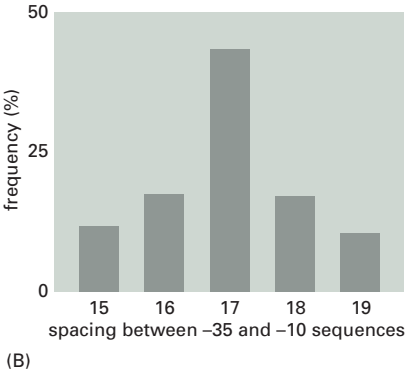


Figure 6–12 Consensus sequence for the major class of *E. coli* promoters. (A) The promoters are characterized by two hexameric DNA sequences, the –35 sequence and the –10 sequence named for their approximate location relative to the start point of transcription (designated +1). For convenience, the 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.

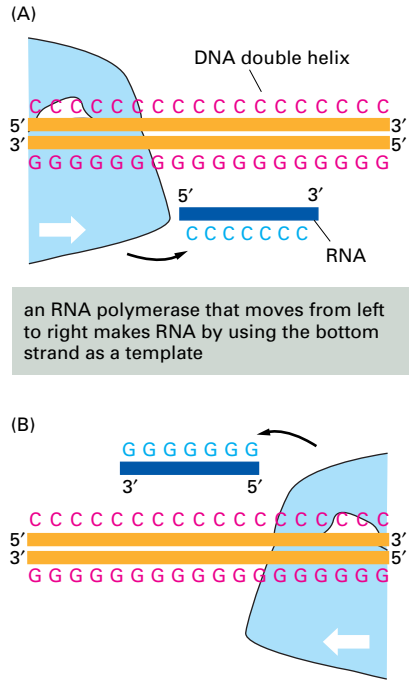
close to form an exit tunnel through which the newly made RNA leaves the enzyme. With the polymerase now functioning in its elongation mode, a rudder-like structure in the enzyme continuously pries apart the DNA–RNA hybrid formed. We can view the series of conformational changes that takes place during transcription initiation as 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 signals in the DNA (termination signals) 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 6–10). As the polymerase transcribes across a terminator, the hairpin may help to wedge open the movable flap on the RNA polymerase and release the RNA transcript from the exit tunnel. At the same time, the DNA–RNA hybrid in the active site, which is held together 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 sufficiently strong enough to hold the RNA in place, and it dissociates causing the release of the polymerase from the DNA, perhaps by forcing open its jaws. 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 ability of RNA to fold into specific structures figures prominantly in many aspects of 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. It is perhaps not surprising that the signals encoded in DNA that specify these transitions are difficult for researchers to recognize. Indeed, a comparison of many different bacterial promoters reveals that they are heterogeneous in DNA sequence. 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). In general, a consensus nucleotide sequence is derived

Figure 6–13 The importance of RNA polymerase orientation. The DNA strand serving as template must be traversed in a 3′ to 5′ direction, as illustrated in Figure 6–9. 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.



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.

One reason that individual bacterial promoters differ in DNA sequence is that the precise sequence determines the strength (or number of initiation events per unit time) of the promoter. Evolutionary processes have thus fine-tuned each promoter to initiate as often as necessary and have 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 include a wide range of sequences, with the potential to form a simple RNA structure being the most important common feature. Since an almost unlimited number of nucleotide sequences have this potential, terminator sequences are much more heterogeneous than those of promoters.

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 develop consensus sequences that summarize their most salient features, their variation in nucleotide sequence makes it difficult for researchers (even when aided by powerful computers) to definitively locate them simply by inspection of the nucleotide sequence of a genome. When we encounter analogous types of sequences in eucaryotes, the problem of locating them is even more difficult. Often, additional information, some of it from direct experimentation, is needed to accurately locate the short DNA signals contained in genomes.

Promoter sequences are asymmetric (see Figure 6–12), and this feature has important consequences for their arrangement 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 nucleotide sequences of bacterial (as well as eucaryotic) promoters are asymmetric the polymerase can bind in only one orientation. The polymerase thus has no option but to transcribe the one DNA strand, since it can synthesize RNA only in the 5′ to 3′ direction (Figure 6–13). The choice of template strand for each gene is therefore determined by the location and orientation of the promoter. Genome sequences reveal that the DNA strand used as the template for RNA synthesis varies from gene to gene (Figure 6–14; see also Figure 1–31).

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.

Transcription Initiation in Eucaryotes Requires Many Proteins

In contrast to bacteria, which contain a single type of RNA polymerase, eucaryotic nuclei have three, called *RNA polymerase I, RNA polymerase II,* and *RNA*

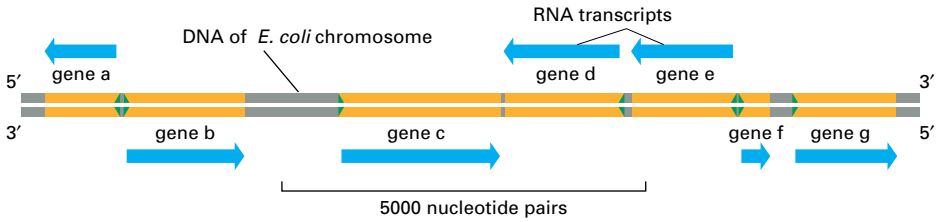


Figure 6–14 Directions of transcription along a short portion of a bacterial chromosome. Some genes are transcribed using one DNA strand as a template, while others are transcribed using the other DNA strand. The direction of transcription is determined by the promoter at the beginning of each gene *(green arrowheads)*. Approximately 0.2% (9000 base pairs) of the *E. coli* chromosome is depicted here. 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 and some snRNA genes
RNA polymerase III	tRNA genes, 5S rRNA genes, some snRNA genes and genes for other small RNAs

polymerase III. The three polymerases are structurally similar to one another (and to the bacterial enzyme). They share some common subunits and many structural features, 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 the vast majority of 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.

1. While bacterial RNA polymerase (with σ factor as one of its subunits) is able to initiate transcription on a DNA template *in vitro* without the help of additional proteins, eucaryotic RNA polymerases cannot. They require the help of a large set of proteins called *general transcription factors*, which must assemble at the promoter with the polymerase before the polymerase can begin transcription.
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 discovery that, unlike bacterial RNA polymerase, purified eucaryotic RNA polymerase II could not initiate transcription *in vitro* led to the discovery and purification of the additional factors required for this process. These **general transcription factors** help to position the 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. The proteins are "general" because they assemble on 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),

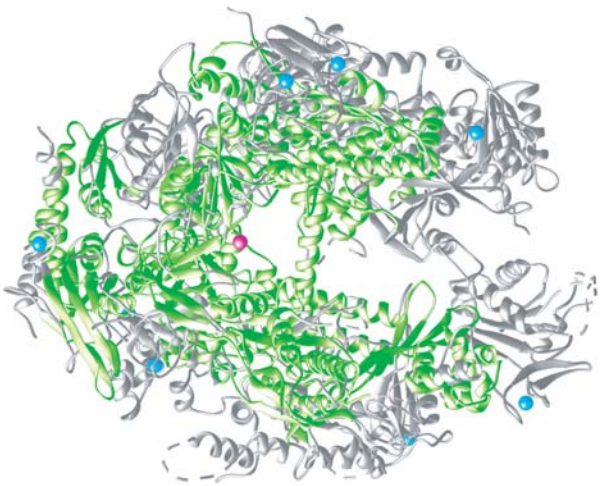


Figure 6–15 Structural similarity between a bacterial RNA polymerase and a eucaryotic RNA polymerase II. Regions of the two RNA polymerases that have similar structures are indicated in *green*. The eucaryotic polymerase is larger than the bacterial enzyme (12 subunits instead of 5), and some of the 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.)

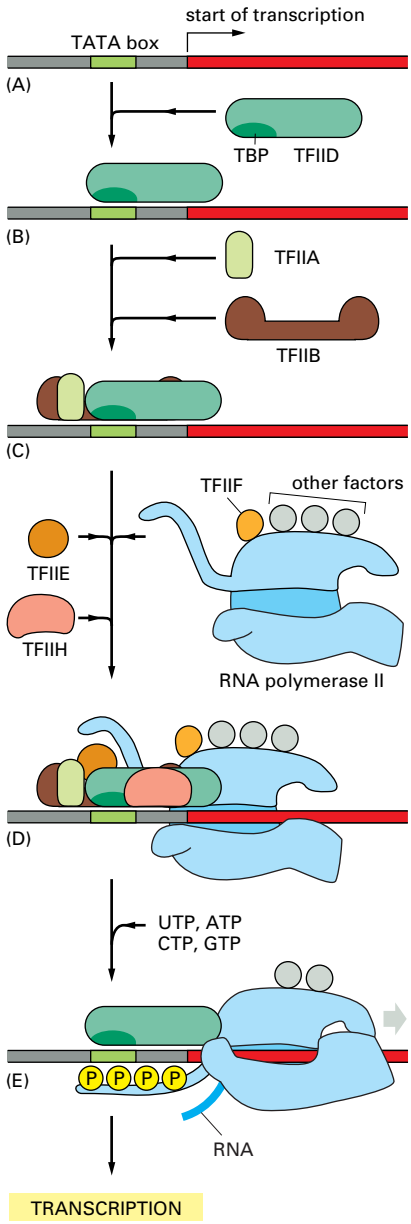
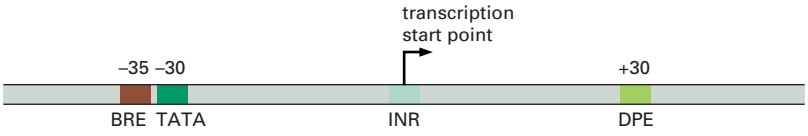


Figure 6–16 Initiation of transcription of a eucaryotic gene by RNA polymerase II. To begin transcription, RNA polymerase requires a number of general transcription factors (called TFIIA, TFIIB, and so on). (A) The promoter contains a DNA sequence called the TATA box, which is located 25 nucleotides away from the site at which transcription is initiated. (B) The TATA box is recognized and bound by transcription factor TFIID, 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, allowing transcription to begin. 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 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 in cells is not known with certainty. In some cases, the general factors are thought to first assemble with the polymerase, with the whole assembly subsequently binding to the DNA in a single step. 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.

and listed as TFIIA, TFIIB, and so on. In a broad sense, the eucaryotic general transcription factors carry out functions equivalent to those of the σ factor in bacteria.

Figure 6–16 shows how the general transcription factors assemble *in vitro* at promoters used by RNA polymerase II. The assembly process starts with the binding of the general transcription factor TFIID 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



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, most polymerase II promoters have a TATA box sequence, and 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.

for most polymerase II promoters, it is the most important. The binding of TFIID causes a large distortion in the DNA 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 are then assembled, along with RNA polymerase II, to form a complete *transcription initiation complex* (see Figure 6–16).

After RNA polymerase II has been guided onto the promoter DNA to form a transcription initiation complex, it must gain access to the template strand at the transcription start point. This step is aided by one of the general transcription factors, TFIIH, which contains a DNA helicase. Next, like the bacterial polymerase, polymerase II remains at the promoter, synthesizing short lengths of RNA until it undergoes a conformational change and is released to begin transcribing a gene. A key step in this release is the addition of phosphate groups to the "tail" of the RNA polymerase (known as the CTD or C-terminal domain). This phosphorylation is also catalyzed by TFIIH, which, in addition to a helicase, contains a protein kinase as one of its subunits (see Figure 6–16, D and E). The polymerase can then disengage from the cluster of general transcription factors, undergoing a series of conformational changes that tighten its interaction with DNA and acquiring new proteins that allow it to transcribe for long distances without dissociating.

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 in position to modify the newly transcribed RNA as it emerges from the polymerase.

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

The model for transcription initiation just described was established by studying the action of RNA polymerase II and its general transcription factors on purified DNA templates *in vitro*. However, as discussed in Chapter 4, DNA in eucaryotic cells is packaged into nucleosomes, which are further arranged in higher-order chromatin structures. As a result, transcription initiation in a eucaryotic cell is more complex and requires more proteins than it does on purified DNA. First, gene regulatory proteins known as *transcriptional activators*

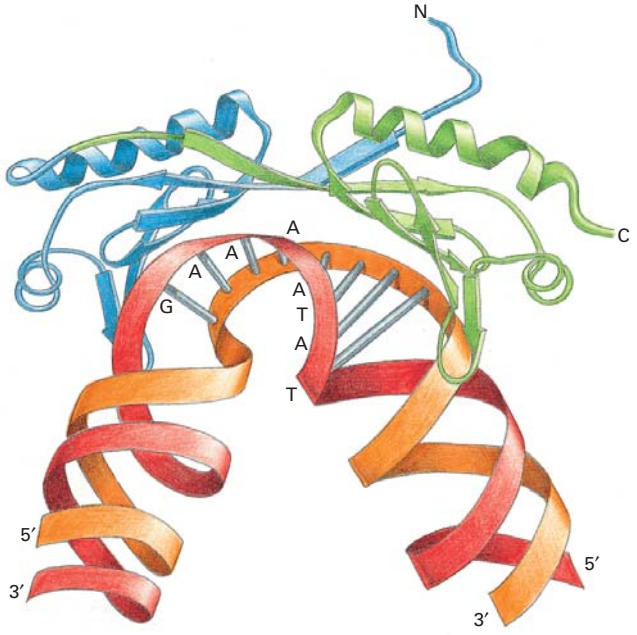


Figure 6–18 Three-dimensional structure of TBP (TATA-binding protein) bound to DNA. The TBP is the subunit of the general transcription factor TFIID that is responsible for recognizing and binding to the TATA box sequence in the DNA *(red)*. The unique DNA bending caused by TBP—two kinks in the double 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.)

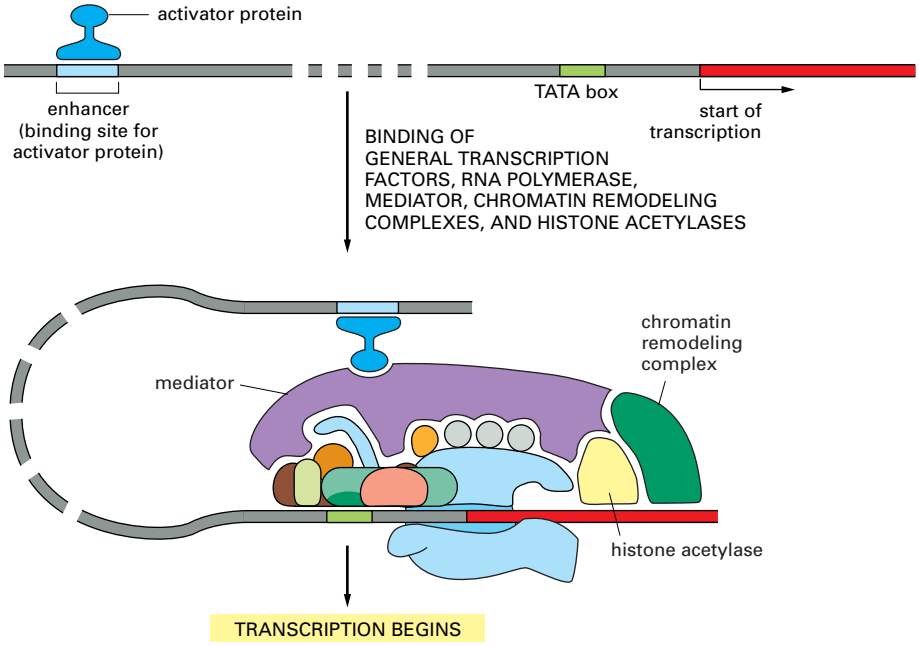


Figure 6–19 Transcription initiation by RNA polymerase II in a eucaryotic cell. Transcription initiation *in vivo* requires the presence of transcriptional activator proteins. As described in Chapter 7, these proteins bind to specific short sequences in DNA. Although only one is shown here, a typical 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 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, and this is likely to be a form of DNA upon which transcription is initiated. For simplicity, it is not shown in the figure.

bind to specific sequences in DNA and help to attract RNA polymerase II to the start point of transcription (Figure 6–19). This attraction is needed to help the RNA polymerase and the general transcription factors in overcoming the difficulty of binding to DNA that is packaged in chromatin. We discuss the role of activators in Chapter 7, because they represent 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 the *mediator*, which allows the activator proteins to communicate properly with the polymerase II and with the general transcription factors. Finally, transcription initiation in the cell often requires the local recruitment of chromatin-modifying enzymes, including chromatin remodeling complexes and histone acetylases (see Figure 6–19). As discussed in Chapter 4, both types of enzymes can allow greater accessibility to the DNA present in chromatin, and by doing so, they facilitate the assembly of the transcription initiation machinery onto DNA.

As illustrated in Figure 6–19, many proteins (well over one hundred individual subunits) must assemble at the start point of transcription to initiate transcription in a eucaryotic cell. The order of assembly of these proteins is probably different for different genes and therefore may not follow a prescribed pathway. In fact, some of these different protein assemblies may interact with each other away from the DNA and be brought to DNA as preformed subcomplexes. For example, the mediator, RNA polymerase II, and some of the general transcription factors can bind to each other in the nucleoplasm and be brought to the DNA as a unit. We return to this issue in Chapter 7, where we discuss the many ways 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 shortly after initiation has occurred and help polymerases to move through the wide

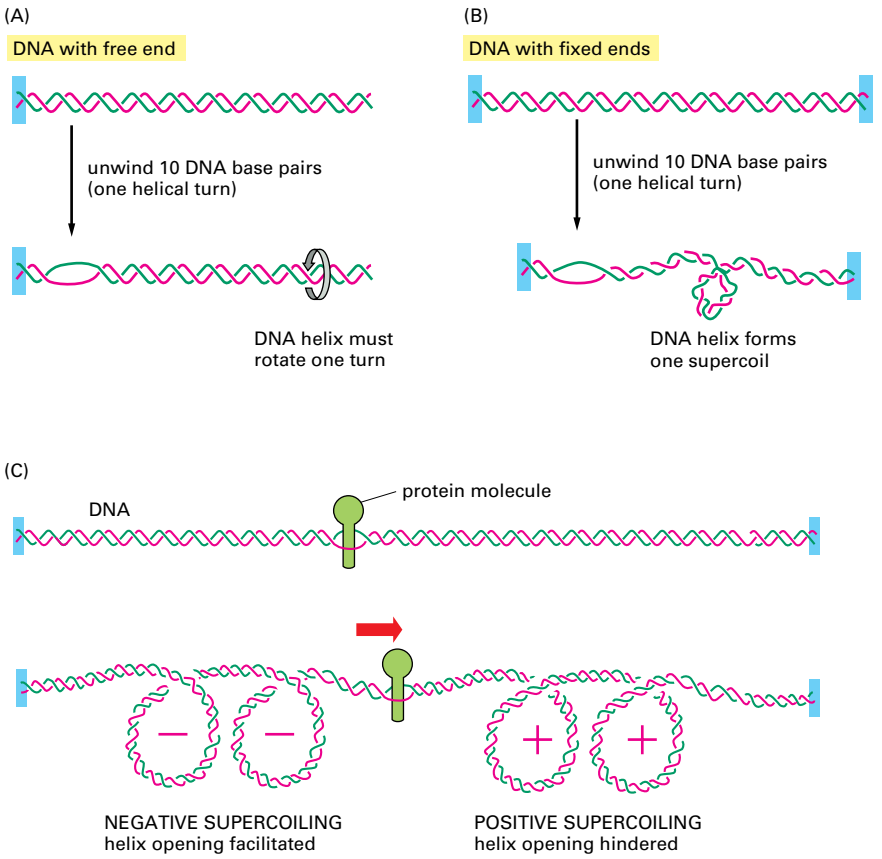


Figure 6–20 Superhelical tension in DNA causes DNA supercoiling. (A) For a DNA molecule with one free end (or a nick in one strand that serves as a swivel), the DNA double helix rotates by one turn for every 10 nucleotide pairs opened. (B) If rotation is prevented, superhelical tension is introduced 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 in this example; 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.

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. Experiments have shown that bacterial polymerases, which never encounter nucleosomes *in vivo*, can nonetheless transcribe through them *in vitro,* suggesting that a nucleosome is easily traversed. However, eucaryotic polymerases have to move through forms of chromatin that are more compact than a simple nucleosome. It therefore seems likely that they transcribe with the aid of chromatin remodeling complexes (see pp. 212–213). 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 how this is accomplished, but these proteins may help to dislodge parts of the nucleosome core as the polymerase transcribes the DNA of a 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 will adopt in response to superhelical tension; conversely, creating various loops or coils in the helix can create such tension. A simple way of visualizing the topological constraints that cause DNA supercoiling is illustrated in Figure 6–20A. 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.

Superhelical tension is also created as RNA polymerase moves along a stretch of DNA that is anchored at its ends (Figure 6–20C). As long as the polymerase is not free to rotate rapidly (and such rotation is unlikely given the size

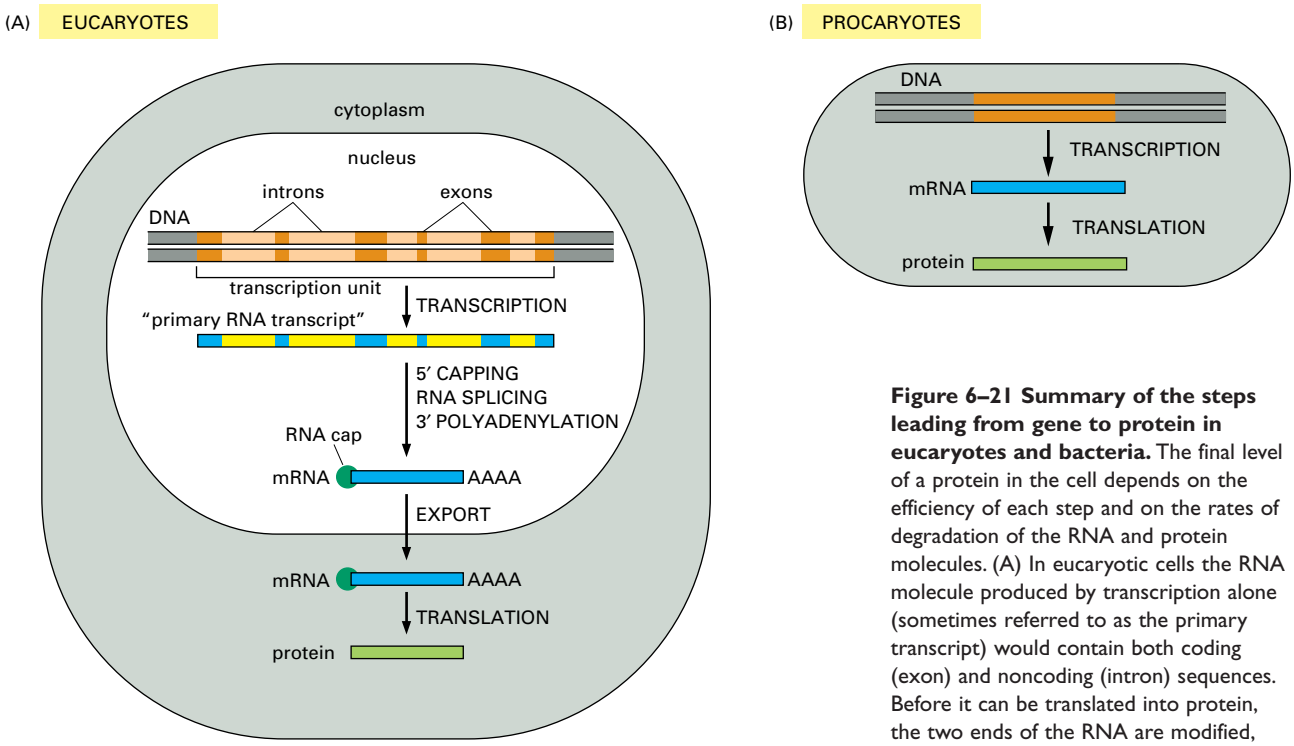


Figure 6–21 Summary of the steps leading from gene to protein in eucaryotes and bacteria. The final level of a protein in the cell depends on the efficiency of each step and on the rates of degradation of the RNA and protein molecules. (A) In eucaryotic cells the RNA molecule produced by transcription alone (sometimes referred to as the primary transcript) would contain 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 these steps are depicted as occurring one at a time, in a sequence, in reality they are coupled and different steps can occur simultaneously. For example, the RNA cap is added and splicing typically begins before transcription has been completed. Because of this coupling, complete primary RNA transcripts do not typically exist in the cell. (B) In procaryotes the production of mRNA molecules is much simpler. The 5′ end of an mRNA molecule is produced by the initiation of transcription by RNA polymerase, 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, translation of a bacterial mRNA often begins before its synthesis has been completed.

of RNA polymerases and their attached transcripts), a moving polymerase generates positive superhelical tension in the DNA in front of it and negative helical tension behind it. For 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. 251). 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). These negative supercoils are removed from bacterial DNA whenever a region of helix opens, 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 6–10).

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 step in a series of reactions that includes the covalent modification of both 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). The modifications of the ends of eucaryotic mRNA are *capping* on the 5′ end and *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 for

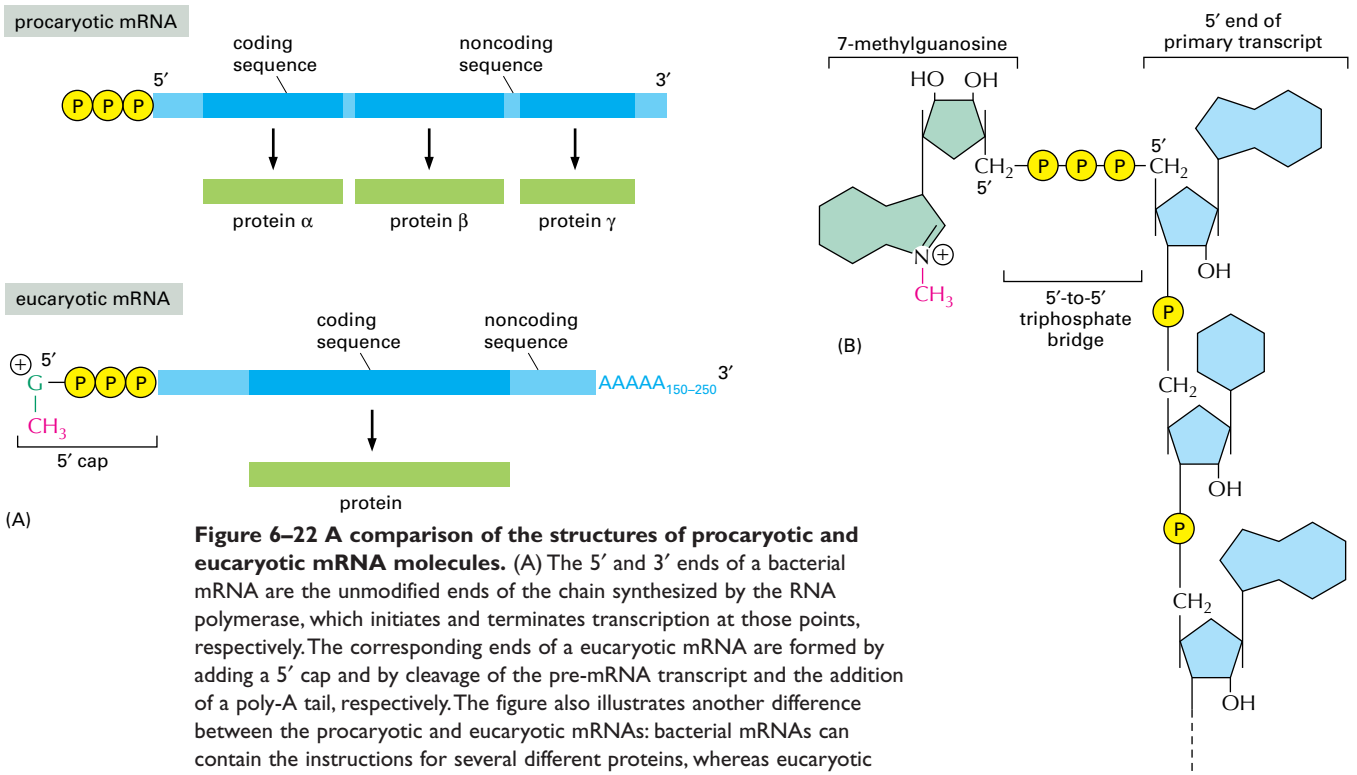


Figure 6–22 A comparison of the structures of procaryotic and eucaryotic mRNA molecules. (A) The 5′ and 3′ ends of a bacterial mRNA are the unmodified ends of the chain synthesized by the RNA polymerase, which initiates and terminates transcription at those points, respectively. The corresponding ends 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).

translation into protein. In Chapter 4, we saw that a typical eucaryotic gene is present in the genome as short blocks of protein-coding sequence (exons) separated by long introns, and RNA splicing is the critically important step in which the different portions of a protein coding sequence are joined together. As we describe next, RNA splicing also provides higher eucaryotes with the ability to synthesize several different proteins from the same gene.

These RNA processing steps are tightly coupled to transcription elongation by an ingenious mechanism. As discussed previously, a key step of the transition of RNA polymerase II to the elongation mode of RNA synthesis is an extensive phosphorylation of the RNA polymerase II tail, called the CTD. This C-terminal domain of the largest subunit consists of a long tandem array of a repeated seven-amino-acid sequence, containing two serines per repeat that can be phosphorylated. Because there are 52 repeats in the CTD of human RNA polymerase II, its complete phosphorylation would add 104 negatively charged phosphate groups to the polymerase. This phosphorylation step not only dissociates the RNA polymerase II from other proteins present at the start point of transcription, it also allows a new set of proteins to associate with the RNA polymerase tail that function in transcription elongation and pre-mRNA 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, RNA polymerase II in its elongation mode can be viewed as an RNA factory that both transcribes DNA into RNA and processes the RNA it produces (Figure 6–23).

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

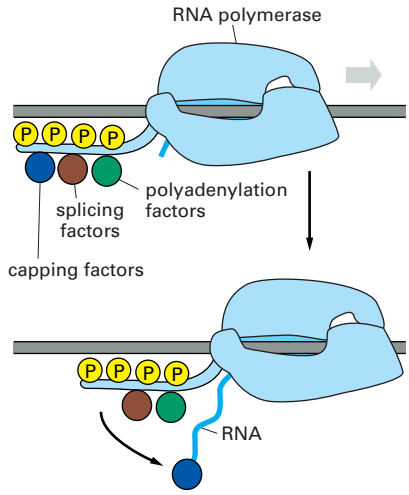


Figure 6–23 The "RNA factory" concept for eucaryotic RNA polymerase II. Not only does the polymerase transcribe DNA into RNA, but it also carries pre-mRNA-processing proteins on its tail, which are then transferred to the nascent RNA at the appropriate time. There are many RNA-processing enzymes, and not all travel with the polymerase. For RNA splicing, for example, only a few critical components are carried on the tail; once transferred to an RNA molecule, they serve as a nucleation site for the remaining components. The RNA-processing proteins first bind to the RNA polymerase tail when it is phosphorylated late in the process of transcription initiation (see Figure 6–16). Once RNA polymerase II finishes transcribing, it is released from DNA, the phosphates on its tail are removed by soluble phosphatases, and it can reinitiate transcription. Only this dephosphorylated form of RNA polymerase II is competent to start RNA synthesis at a promoter.

a modified guanine nucleotide (see Figure 6–22B). The capping reaction is performed by three enzymes acting in succession: one (a phosphatase) removes one 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 phosphorylated RNA polymerase tail, they are poised to modify the 5′ end of the nascent transcript as soon as it emerges from the polymerase.

The 5′-methyl cap signals 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 RNAs during transcription, in part because these polymerases lack tails. 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 type. It is termed precursor-mRNA (or pre-mRNA) splicing to denote that it occurs on RNA molecules destined to become mRNAs. 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 phosphate bonds remains the same, these reactions could in principle take place without nucleoside triphosphate hydrolysis. However, the machinery that catalyzes pre-mRNA splicing is complex, consisting of 5 additional RNA molecules and over 50 proteins, and it hydrolyzes many ATP molecules per splicing event. This complexity is presumably needed to ensure that splicing is highly accurate, while also being sufficiently flexible to deal with the enormous variety of introns found in a typical eucaryotic cell. Frequent mistakes in RNA

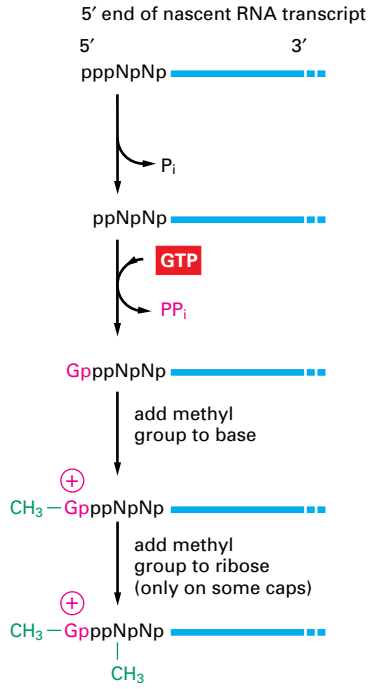


Figure 6–24 The reactions that cap the 5′ end of each RNA molecule synthesized by RNA polymerase II. The final cap contains a novel 5′-to-5′ linkage between the positively charged 7-methyl G residue and the 5′ end of the RNA transcript (see Figure 6–22B). The letter N represents any one of the four ribonucleotides, 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. © ICSU Press.)

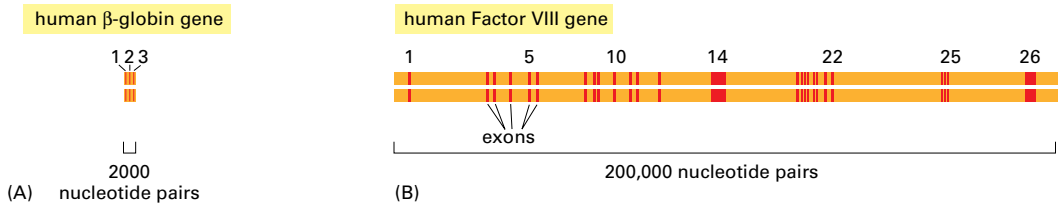


Figure 6–25 Structure of two human genes showing the arrangement of exons and introns. (A) The relatively small β-globin gene, which encodes one of the subunits of the oxygen-carrying protein hemoglobin, contains 3 exons (see also Figure 4–7). (B) The much larger Factor VIII gene contains 26 exons; it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

splicing would severely harm the cell, as they would result in malfunctioning proteins. We see in Chapter 7 that when rare splicing mistakes do occur, the cell has a "fail-safe" device to eliminate the incorrectly spliced mRNAs.

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. Thus, the presence of numerous introns in DNA allows genetic recombination to readily combine the exons of different genes (see p. 462), allowing genes for new proteins to evolve more easily by the combination of parts of preexisting genes. This idea is supported by the observation, described in Chapter 3, that many proteins in present-day cells resemble patchworks composed from a common set of protein pieces, called protein *domains*.

RNA splicing also has a present-day advantage. The transcripts of many eucaryotic genes (estimated at 60% of genes in humans) are spliced in a variety of different ways to produce a set of different mRNAs, thereby allowing a corresponding set of different proteins to be produced from the same gene (Figure 6–27). We discuss additional examples of alternative splicing in Chapter 7, as this is also one of the mechanisms that cells use to change expression of their genes. 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 several times in this chapter and the next, but we first need to describe the cellular machinery that performs this remarkable task.

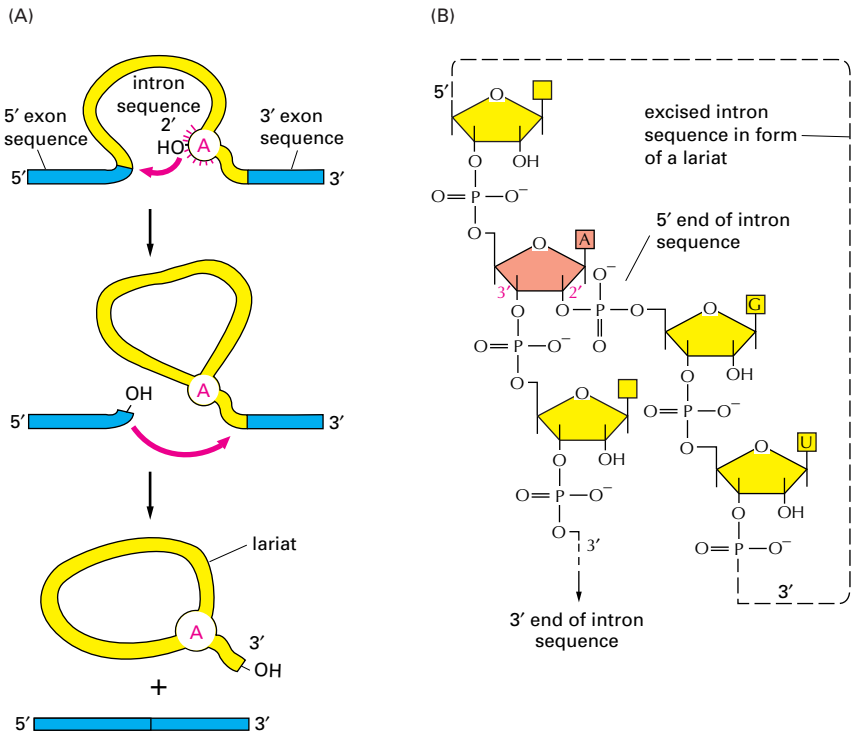


Figure 6–26 The RNA splicing reaction. (A) In the first step, a specific adenine nucleotide in the intron sequence (indicated in *red*) attacks the 5′ splice site and cuts the sugar-phosphate backbone of the RNA at this point. The cut 5′ end of the intron becomes covalently linked to the adenine nucleotide, 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 degraded in due course.

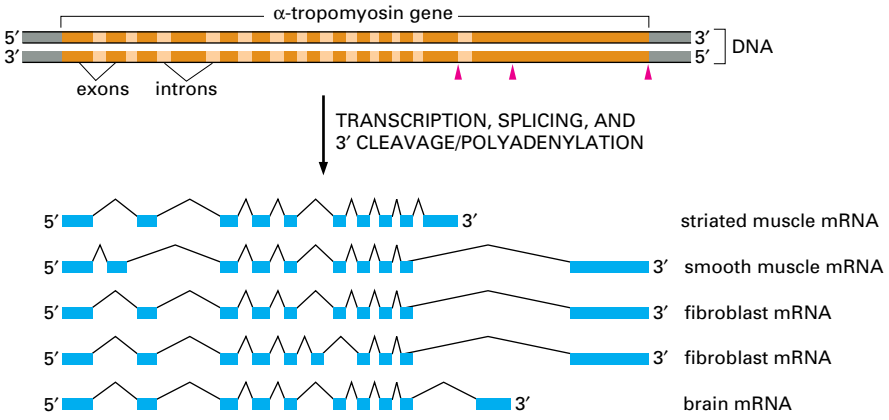


Figure 6–27 Alternative splicing of the α-tropomyosin gene from rat. α-tropomyosin is a coiled-coil protein (see Figure 3–11) that regulates contraction in muscle cells. The primary transcript can be spliced in different ways, as indicated in the figure, to produce distinct mRNAs, which then give rise 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 demark the sites where cleavage and poly-A addition can occur.

Nucleotide Sequences Signal Where Splicing Occurs

Introns range in size from about 10 nucleotides to over 100,000 nucleotides. Picking out the precise borders of an intron is very difficult for scientists to do (even with the aid of computers) when confronted by a complete genome sequence of a eucaryote. The possibility of alternative splicing compounds the problem of predicting protein sequences solely from a genome sequence. This difficulty constitutes one of the main barriers to identifying all of the genes in a complete genome sequence, and it is the primary reason that we know only the approximate number of genes in, for example, the human genome. Yet each cell in our body recognizes and rapidly excises the appropriate intron sequences with high fidelity. We have seen that intron sequence removal involves three positions on the RNA: the 5′ splice site, the 3′ splice site, and the branch point in the intron sequence that forms the base of the excised lariat. In pre-mRNA splicing, each of these three sites has a consensus nucleotide sequence that is similar from intron to intron, providing the cell with cues on where splicing is to take place (Figure 6–28). However, there is enough variation in each sequence to make it very difficult for scientists to pick out all of the many splicing signals in a genome sequence.

RNA Splicing Is Performed by the Spliceosome

Unlike the other steps of mRNA production we have discussed, RNA splicing is performed largely by RNA molecules instead of proteins. RNA molecules recognize intron–exon borders and 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 **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.

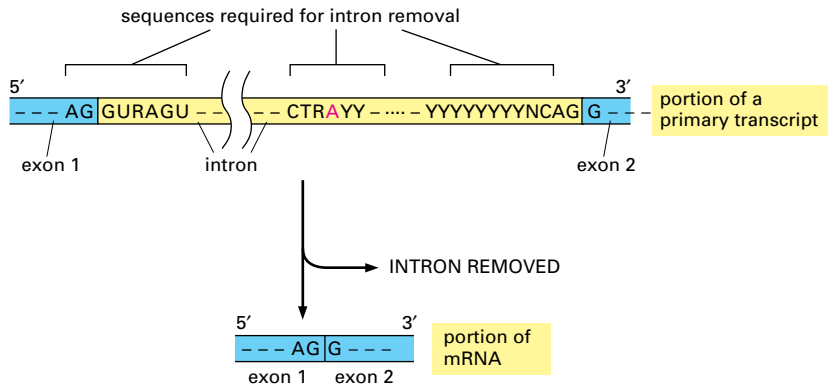


Figure 6–28 The consensus nucleotide sequences in an RNA molecule that signal the beginning and the end of most introns in humans. Only the three blocks of nucleotide sequences shown are required to remove an intron sequence; the rest of the intron can be occupied by any nucleotide. Here A, G, U, and C are the standard RNA nucleotides; R stands for either A or G; Y stands for either 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. The remaining positions (even the branch point A) can be occupied by a variety of nucleotides, 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.

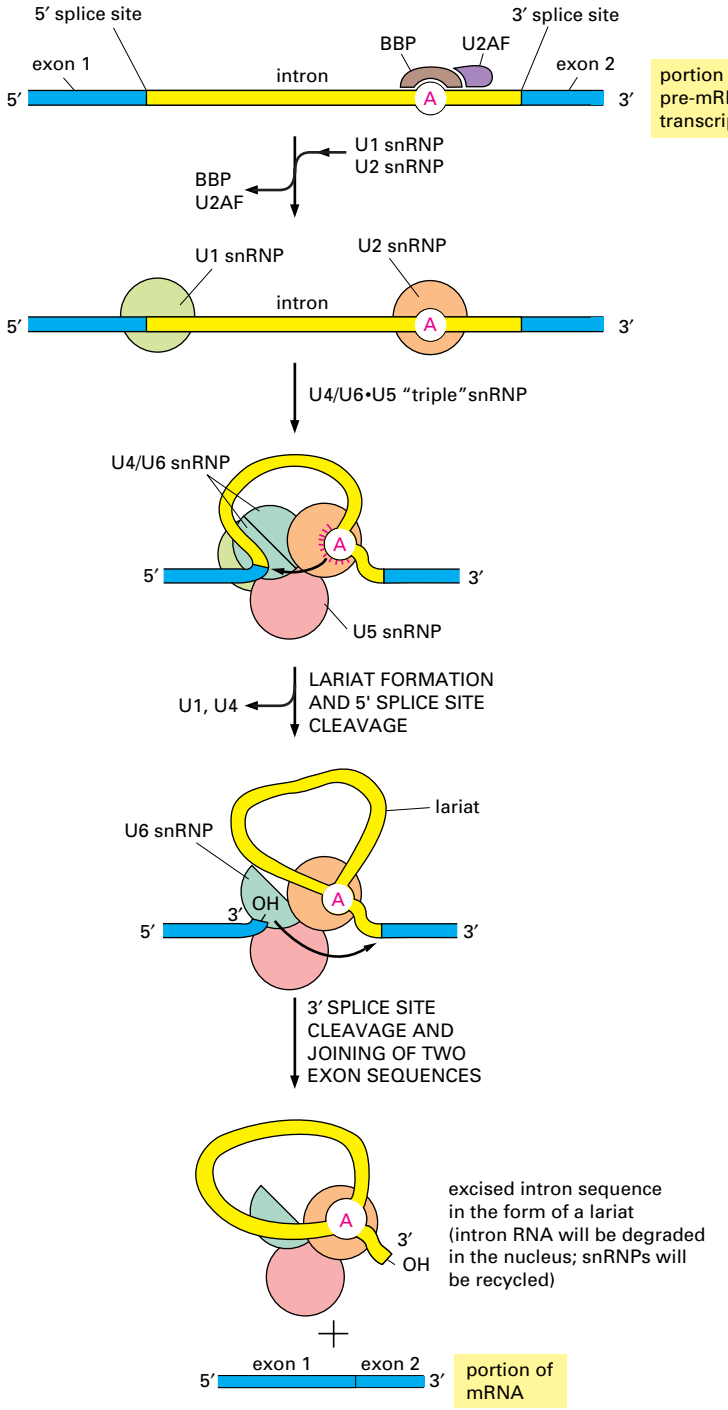


Figure 6–29 The RNA splicing mechanism. RNA splicing is catalyzed by an assembly of snRNPs (shown as *colored circles)* plus other proteins (most of which are not shown), which together constitute the spliceosome. The spliceosome recognizes the splicing signals on a pre-mRNA molecule, brings the two ends of the intron together, and provides the enzymatic activity for the two reaction steps (see Figure 6–26). The branch-point site is first recognized by the BBP (branch-point binding protein) and U2AF, a helper protein. In the next steps, the U2 snRNP displaces BBP and U2AF and forms base pairs with the branch-point site consensus sequence, and the U1 snRNP forms base-pairs with the 5′ splice junction (see Figure 6–30). At this point, the U4/U6•U5 "triple" snRNP enters the spliceosome. In this triple snRNP, the U4 and U6 snRNAs are held firmly together by base-pair interactions and the U5 snRNP is more loosely associated. Several RNA–RNA rearrangements then occur that break apart the U4/U6 base pairs (as shown, the U4 snRNP is ejected from the splicesome before splicing is complete) and allow the U6 snRNP to displace U1 at the 5′ splice junction (see Figure 6–30). Subsequent rearrangements create the active site of the spliceosome and position the appropriate portions of the pre-mRNA substrate for the splicing reaction to occur. Although not shown in the figure, each splicing event requires additional proteins, some of which hydrolyze ATP and promote the RNA–RNA rearrangements.

The spliceosome is a dynamic machine; as we see below, it is assembled on pre-mRNA from separate components, and parts enter and leave it as the splicing reaction proceeds (Figure 6–29). 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 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). As we shall see, 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 stepwise assembly and rearrangements of the spliceosome. Some of the additional proteins that make up the spliceosome are RNA helicases, which use the energy of ATP hydrolysis to break existing RNA–RNA interactions so as 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. In all, more than 50 proteins, including those that form the snRNPs, are required for each splicing event.

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 roles 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 an important way of preventing wayward splicing.

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.

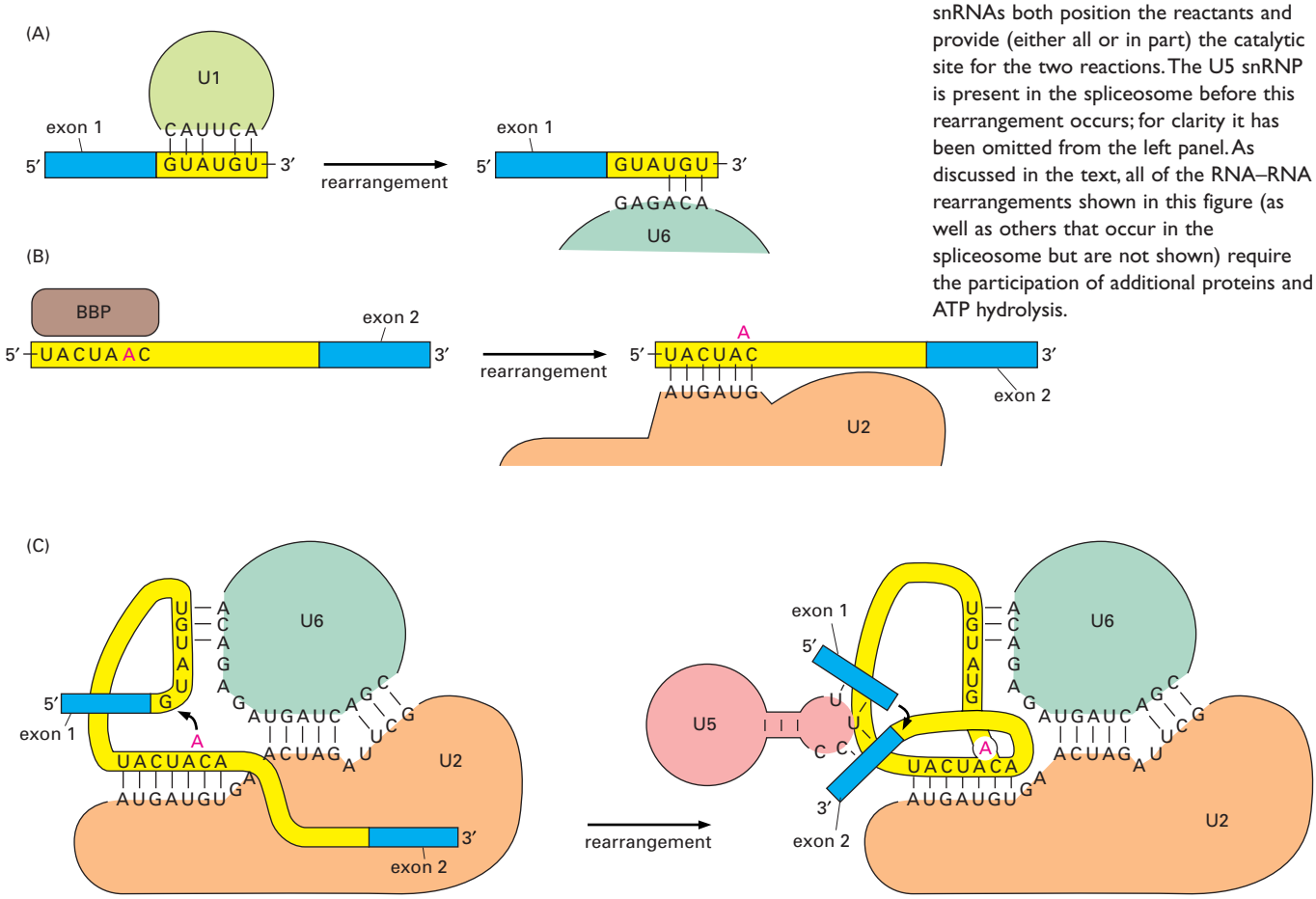


Figure 6–30 Several of the rearrangements that take place in the spliceosome during pre-mRNA splicing. Shown here are the details for the yeast *Saccharomyces cerevisiae*, in which the nucleotide sequences involved are slightly different from those in human cells. (A) The exchange of U1 snRNP for U6 snRNP occurs before the first phosphoryl-transfer reaction (see Figure 6–29). This exchange allows 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, U5 snRNP undergoes a rearrangement that 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 site 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.

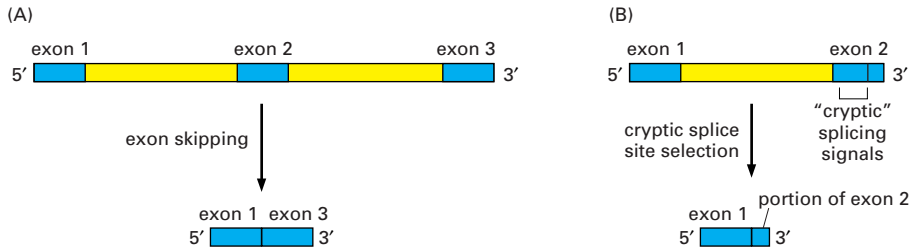


Figure 6–31 Two types of splicing errors. Both types might be expected to occur frequently if splice-site selection were performed by the spliceosome on a preformed, protein-free RNA molecule. "Cryptic" splicing signals are nucleotide sequences of RNA that closely resemble true splicing signals.

Once the splicing chemistry is completed, the snRNPs remain bound to the lariat and the spliced product is released. 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.

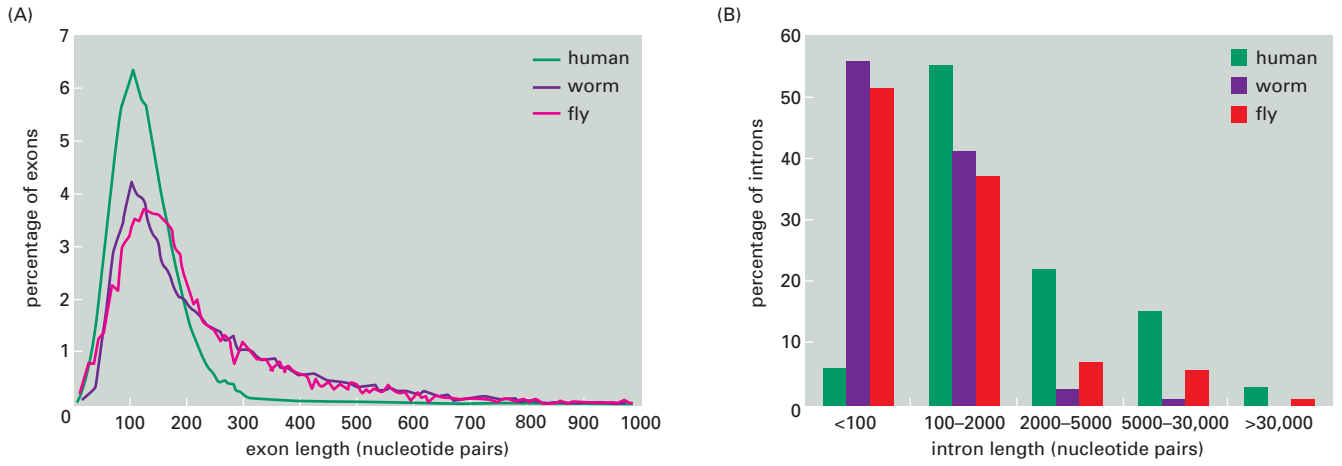
Ordering Influences in the Pre-mRNA Help to Explain How the Proper Splice Sites Are Chosen

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 are supplemented by two additional factors that help ensure that splicing occurs accurately. These ordering influences in the pre-mRNA increase the probability that the appropriate pairs of 5′ and 3′ splice sites will be brought together in the spliceosome before the splicing chemistry begins. The first results from the assembly of the spliceosome occurring as the pre-mRNA emerges from a transcribing RNA polymerase II (see Figure 6–23). As for 5′ cap formation, several components of the spliceosome seem to be carried on the phosphorylated tail of RNA polymerase. Their transfer directly from the polymerase to the nascent pre-mRNA presumably helps the cell to keep track of introns and exons: the snRNPs 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. This feature helps to prevent inappropriate exon skipping.

The second factor that helps the cell to choose splice sites has been termed the "exon definition hypothesis," and it is understood only in outline. 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). As RNA synthesis proceeds, a group of spliceosome components, called the SR proteins (so-named because they contain a domain rich in serines and arginines), are thought to assemble on exon sequences and mark off each 3′ and 5′ splice site starting at the 5′ end of the RNA (Figure 6–33). This assembly takes place in conjunction with the U1 snRNA, which marks one exon boundary, and U2AF,

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.)



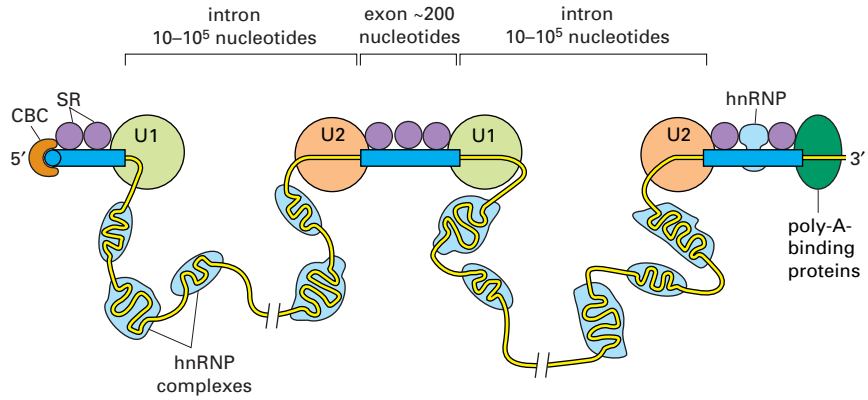


Figure 6–33 The exon definition hypothesis. According to one proposal, SR proteins bind to each exon sequence in the pre-mRNA and thereby help to guide the snRNPs to the proper intron/exon boundaries. This demarcation of exons by the SR proteins occurs co-transcriptionally, beginning at the CBC (cap-binding 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. Each hnRNP complex forms a particle approximately twice the diameter of a nucleosome, and the core is composed of a set of at least eight different proteins. It has been proposed that hnRNP proteins preferentially associate with intron sequences and that this preference also helps the spliceosome distinguish introns from exons. However, as shown, at least some hnRNP proteins may bind to exon sequences but their role, if any, in exon definition has yet to be established. (Adapted from R. Reed, *Curr. Opin. Cell Biol.* 12:340–345, 2000.)

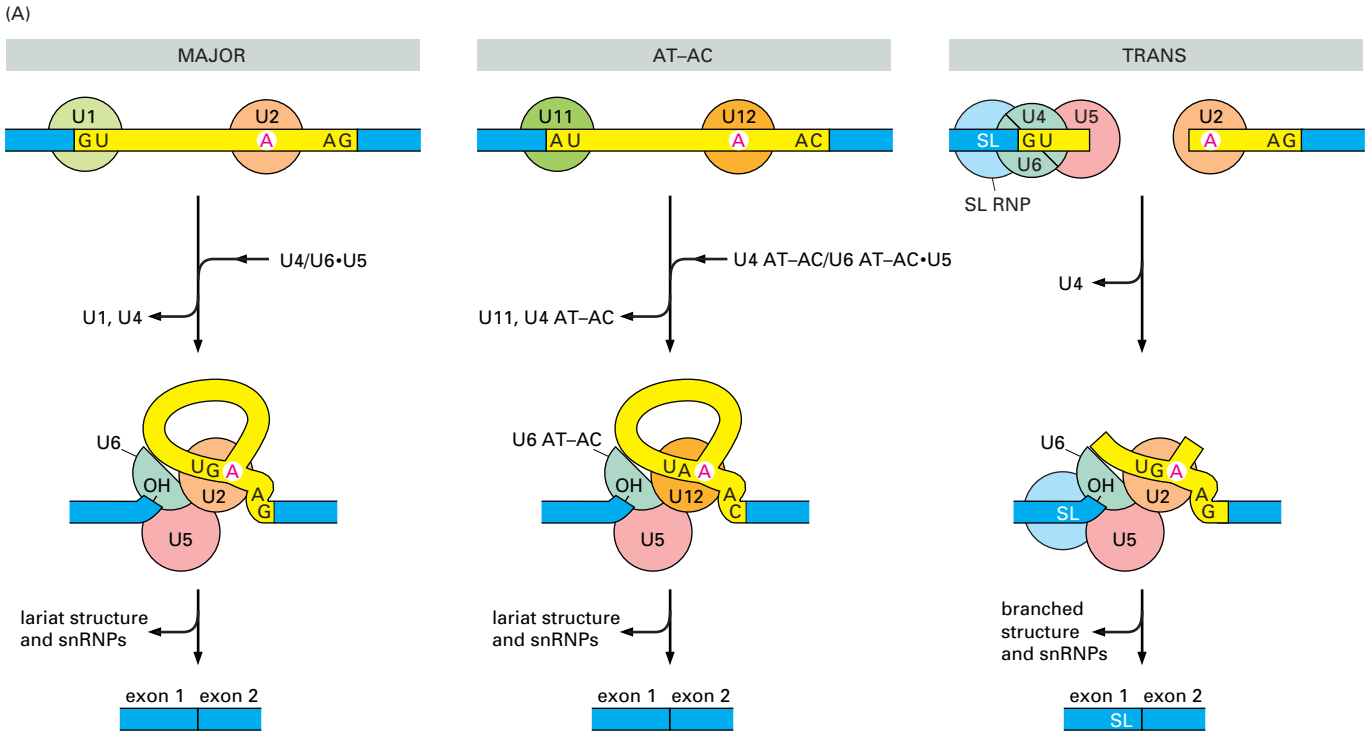
which initially helps to specify the other. By specifically marking the exons in this way, the cell increases the accuracy with which the initial splicing components are deposited on the nascent RNA and thereby helps to avoid cryptic splice sites. How the SR proteins discriminate exon sequences from intron sequences is not understood; however, it is known that some of the SR proteins bind preferentially to RNA sequences in specific exons. In principle, the redundancy in the genetic code could have been exploited during evolution to select for binding sites for SR proteins in exons, allowing these sites to be created without constraining amino acid sequences.

Both the marking out 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 yeast 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 DNA sequences at the 5′ and 3′ splice junctions and at the branch point; it is called the *AT–AC spliceosome* because of the nucleotide sequence determinants at its intron–exon borders (Figure 6–34). 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). The recent discovery of this class of snRNPs gives us confidence in the base-pair interactions deduced for the major spliceosome, because it provides an independent set of molecules that undergo the same RNA–RNA interactions despite differences in the RNA sequences involved.

A particular variation on splicing, called **trans-splicing**, has been discovered in a few eucaryotic organisms. These 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 6–34). Trypanosomes produce all of their mRNAs in this way, whereas only about 1% of nematode mRNAs are produced by trans-splicing. 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).



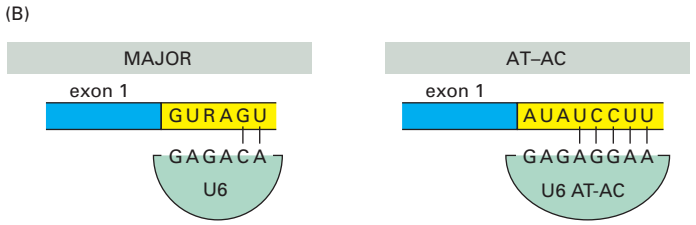


Figure 6–34 Outline of the mechanisms used for three types of RNA splicing. (A) Three types of spliceosomes. The major spliceosome *(left)*, the AT–AC spliceosome (middle), and the trans-spliceosome *(right)* are each shown at two stages of assembly. The U5 snRNP is the only component that is common to all three spliceosomes. Introns removed by the AT–AC 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 AT–AC spliceosome. In trans-splicing, 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 AT–AC snRNP 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.)

The reason that a few organisms use trans-splicing is not known; however, it is thought that the common 5′ exon may aid in the translation of the mRNA. Thus, the products of 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 many features of the pre-mRNA transcript; these include the affinity of the three signals on the RNA (the 5′ and 3′ splice junctions and branch point) for the splicing machinery, the length and nucleotide sequence of the exon, the co-transcriptional assembly of the spliceosome, and the accuracy of the "bookkeeping" that underlies exon definition. So far we have emphasized the accuracy of the RNA splicing processes that occur in a cell. But it also seems that the mechanism has been selected for its flexibility, which allows the cell to try out new proteins on occasion. Thus, for example, when a mutation occurs in a nucleotide sequence critical for splicing of a particular intron, it 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 used (Figure 6–35C). Presumably, 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 easily regulate the pattern of RNA splicing. Earlier in this section we saw that alternative splicing

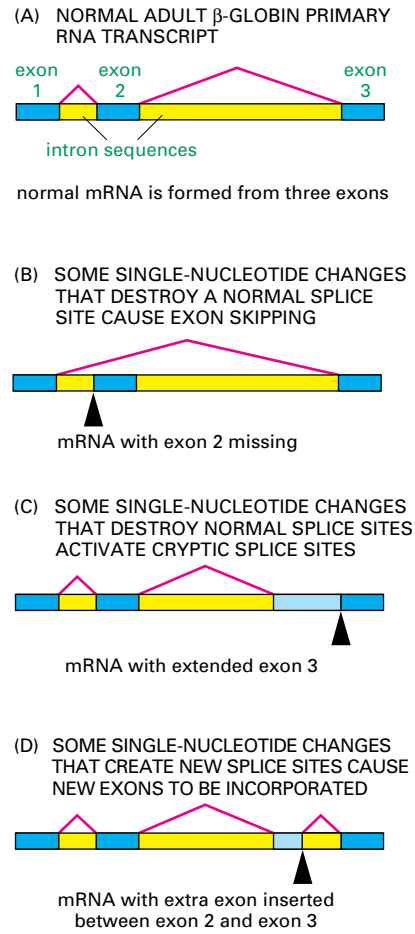


Figure 6–35 Abnormal processing of the β-globin primary RNA transcript in humans with the disease β thalassemia. In the examples shown, the disease is caused by splice-site mutations, denoted by *black arrowheads*. The *dark blue boxes* represent the three normal exon sequences; the *red lines* are used to indicate the 5′ and 3′ splice sites that are used in splicing the RNA transcript. 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.)

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 most cases, the splicing patterns are regulated by the cell 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 taken up again in more detail in the final section of this chapter), it is thought 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). Based on the fact that 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 II 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.

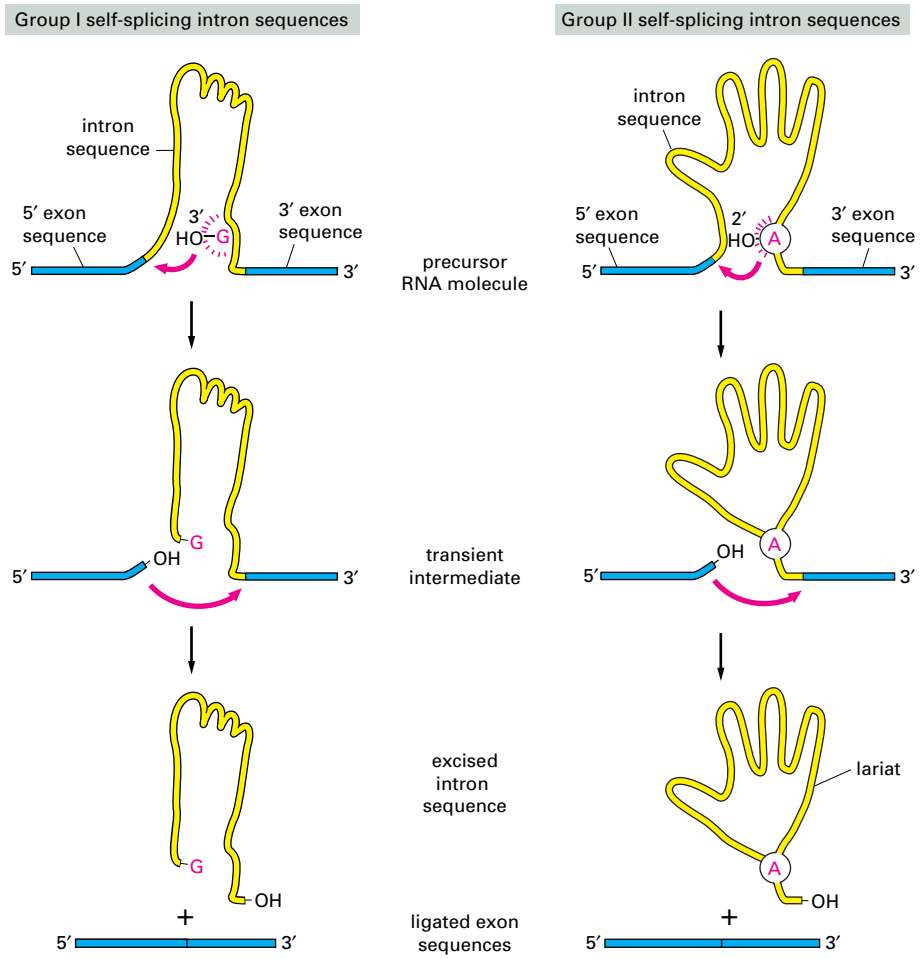


Figure 6–36 The two known classes of self-splicing intron sequences. The group I intron sequences bind a free G nucleotide to a specific site on the RNA to initiate splicing, while the group II intron sequences use an especially reactive A nucleotide in the intron sequence itself for the same purpose. The two mechanisms have been drawn to emphasize their similarities. Both are normally aided in the cell by proteins that speed up the reaction, but the catalysis is nevertheless mediated by the RNA in the intron sequence. 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 great majority of RNA splicing in eucaryotic cells is performed by the spliceosome, and self-splicing RNAs represent unusual cases. (Adapted from T.R. Cech, *Cell* 44:207–210, 1986.)

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 components assemble on the RNA and delineate 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 the RNA emerges from the enzyme. As we see in this section, as RNA polymerase II terminates transcription at the end of a gene, it uses a similar mechanism to ensure that the 3′ end of the pre-mRNA becomes appropriately processed.

As might be expected, the 3′ ends of mRNAs are ultimately specified by DNA signals encoded in the genome (Figure 6–37). These DNA signals are transcribed

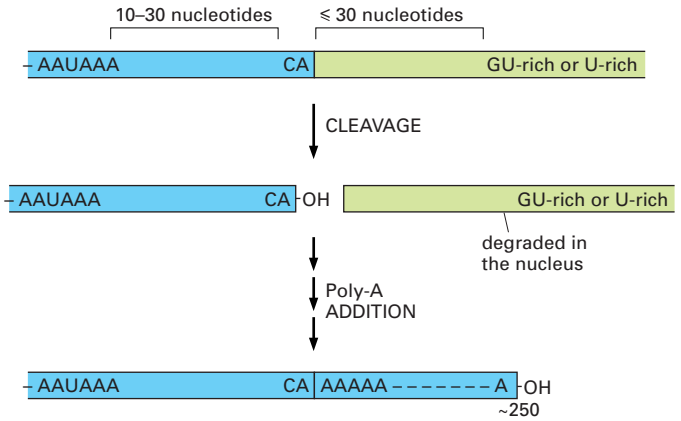


Figure 6–37 Consensus nucleotide sequences that direct cleavage and polyadenylation to form the 3′ end of a eucaryotic mRNA. These sequences are encoded in the genome and are recognized by specific proteins after they are transcribed into RNA. The hexamer AAUAAA is bound by CPSF, the GU-rich element beyond the cleavage site by CstF (see Figure 6–38), and the CA sequence 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.

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 F) 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. Some of the subunits of CPSF are associated with the general transcription factor TFIID, which, as we saw earlier in this chapter, is involved in transcription initiation. During transcription initiation, these subunits may be transferred from TFIID to the RNA polymerase tail, remaining associated there until the polymerase has transcribed through the end of a gene.

Once CstF and CPSF bind to specific nucleotide sequences on an emerging RNA molecule, additional proteins assemble with them to perform the processing that creates the 3′ end of the mRNA. First, the RNA is cleaved (see Figure 6–38). Next an enzyme called poly-A polymerase 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 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. Poly-A-binding proteins remain bound to the poly-A tail as the mRNA makes its journey 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 continuing as many as several hundred nucleotides beyond the DNA that contains the 3′ cleavage-site information. But the polymerase soon releases its grip on the template and transcription terminates; the piece of RNA downstream of the cleavage site is then degraded in the cell nucleus. It is not yet understood what triggers the loss in polymerase II processivity after the RNA is cleaved. One idea is that the transfer of the 3′ end processing factors from the RNA polymerase to the RNA causes a conformational change in the polymerase that loosens its hold on DNA; another is that the lack of a cap structure (and the CBC) on the 5′ end of the RNA that emerges from the polymerase somehow signals to the polymerase to terminate transcription.

Mature Eucaryotic mRNAs Are Selectively Exported from the Nucleus

We have seen how eucaryotic pre-mRNA synthesis and processing takes 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 spliced pre-mRNAs—is not only useless but could be dangerous if it was not destroyed. 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 transport of mRNA from the nucleus to the cytoplasm, where it is translated into protein, is highly selective—being closely coupled to correct RNA processing. This coupling is achieved by the *nuclear pore complex,* which recognizes and transports only completed mRNAs.

We have seen that as a pre-mRNA molecule is synthesized and processed, it is bound by a variety of proteins, including the cap-binding complex, the SR proteins, and the poly-A binding proteins. To be "export-ready," it seems than an mRNA must be bound by the appropriate set of proteins—with certain proteins such as the cap-binding complex being present, and others such as snRNP proteins absent. Additional proteins, placed on the RNA during splicing, seem to

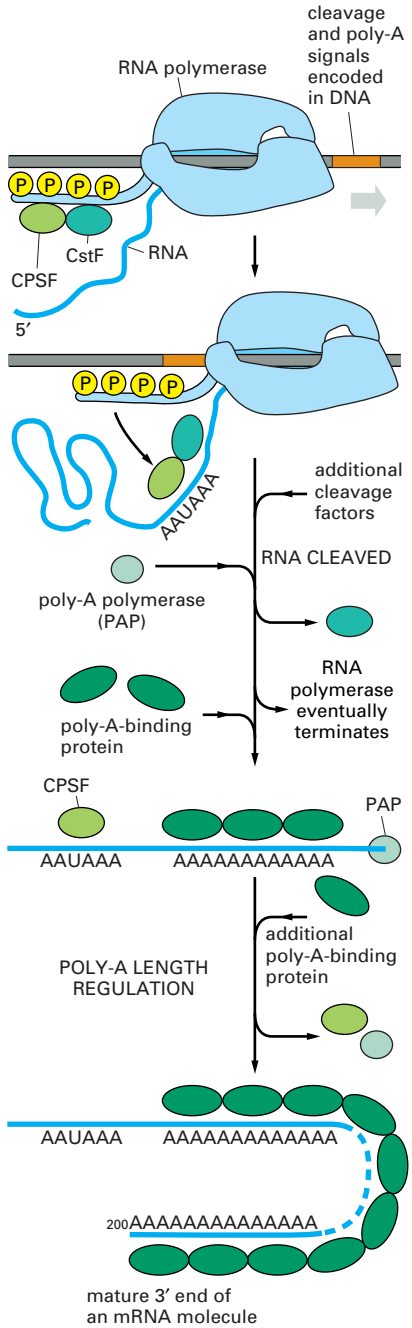


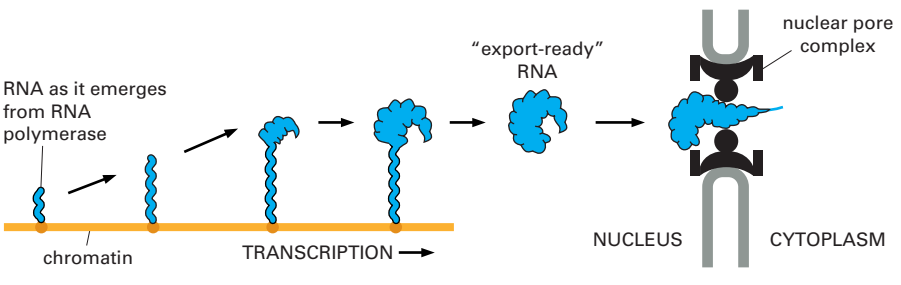
Figure 6–38 Some of the major steps in generating the 3′ end of a eucaryotic mRNA. This process is much more complicated than the analogous process in bacteria, where the RNA polymerase simply stops at a termination signal and releases both the 3′ end of its transcript and the DNA template (see Figure 6–10).

mark exon-exon boundaries and thereby signify completed splicing events. Only if the proper set of proteins is bound to an mRNA is it guided through the **nuclear pore complex** into the cytosol. As described in Chapter 12, nuclear pore complexes are aqueous channels in the nuclear membrane that directly connect the nucleoplasm and cytosol. Small molecules (less than 50,000 daltons) can diffuse freely through them. However, most of the macromolecules in cells, including mRNAs complexed with proteins, are far too large to pass through the pores without a special process to move them. An active transport of substances through the nuclear pore complexes occurs in both directions. As explained in Chapter 12, signals on the macromolecule determine whether it is exported from the nucleus (a mRNA, for example) or imported into it (an RNA polymerase, for example). For the case of mRNAs, the bound proteins that mark completed splicing events are of particular importance, as they are known to serve directly as RNA export factors (see Figure 12–16). mRNAs transcribed from genes that lack introns apparently contain nucleotide sequences that are directly recognized by other RNA export factors. Eucaryotic cells thus use their nuclear pore complexes as gates that allow only useful RNA molecules to enter the cytoplasm.

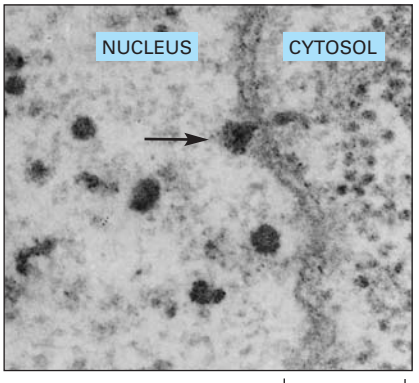
Of all the proteins that assemble on pre-mRNA molecules as they emerge from transcribing RNA polymerases, the most abundant are the hnRNPs (heterogeneous nuclear ribonuclear proteins). Some of these proteins (there are approximately 30 of them in humans) remove the hairpin helices from the RNA so that splicing and other signals on the RNA can be read more easily. Others package the RNA contained in the very long intron sequences typically found in genes of complex organisms (see Figure 6–33). Apart from histones, certain hnRNP proteins are the most abundant proteins in the cell nucleus, and they may play a particularly important role in distinguishing mature mRNA from processing debris. hnRNP particles (nucleosome-like complexes of hnRNP proteins and RNA—see Figure 6–33) are largely excluded from exon sequences, perhaps by prior binding of spliceosome components. They remain on excised introns and probably help mark them for nuclear retention and eventual destruction.

The export of mRNA–protein complexes from the nucleus can be observed with an 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 hnRNP and SR proteins). This protein–RNA complex undergoes a series of structural transitions, probably reflecting RNA processing events, culminating in a curved fiber (Figure 6–39). This curved fiber then moves through the nucleoplasm and enters the nuclear pore complex (with its 5′ cap proceeding first), and it undergoes another series of structural transitions as it moves through the NPC. 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, export, and translation (Figure 6–40).

Before discussing what happens to mRNAs after they leave the nucleus, we briefly consider how the synthesis and processing of noncoding RNA molecules



(A)



(B)

Figure 6–39 Transport of a large mRNA molecule through the nuclear pore complex. (A) The maturation of a Balbiani Ring mRNA molecule as it is synthesized by RNA polymerase and packaged by a variety of nuclear proteins. This drawing of unusually abundant RNA produced by an insect cell is based on EM micrographs such as that shown in (B). Balbiani Rings are described in Chapter 4. (A, adapted from B. Daneholt, *Cell* 88:585–588, 1997; B, from B.J. Stevens and H. Swift, *J. Cell Biol.* 31:55–77, 1966. © The Rockefeller University Press.)

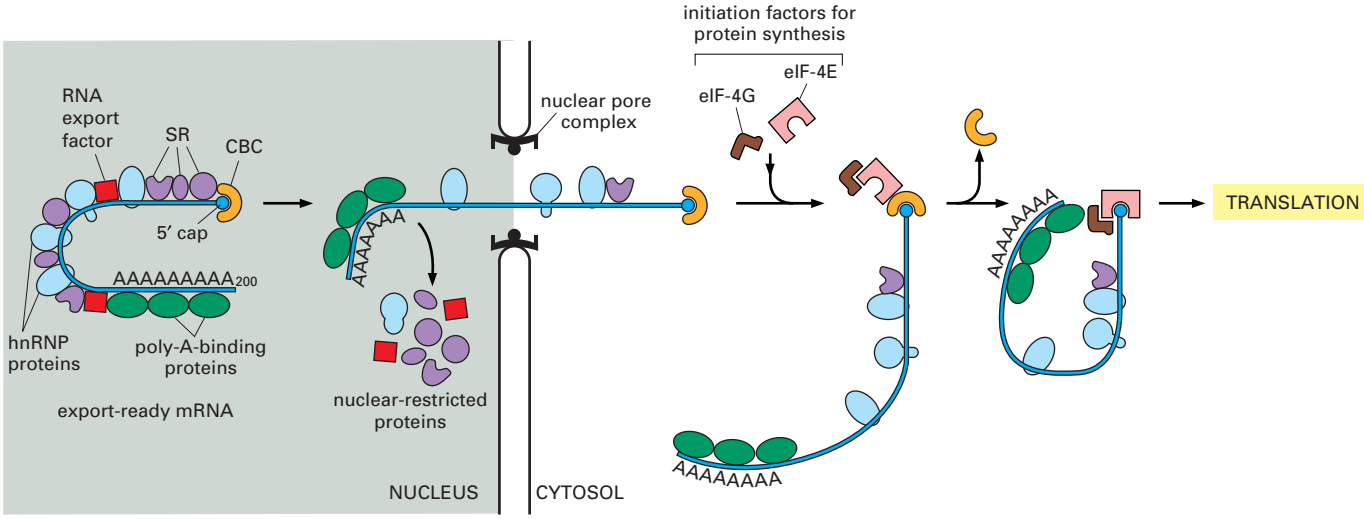


Figure 6–40 Schematic illustration of an "export-ready" mRNA molecule and its transport through the nuclear pore. As indicated, some proteins travel with the mRNA as it moves through the pore, whereas others remain in the nucleus. Once in the cytoplasm, the mRNA continues to shed previously bound proteins and acquire new ones; these substitutions affect the subsequent translation of the message. Because some are transported with the RNA, the proteins that become bound to an mRNA in the nucleus can influence its subsequent stability and translation in the cytosol. RNA export factors, shown in the nucleus, play an active role in transporting the mRNA to the cytosol (see Figure 12–16). Some are deposited at exon-exon boundaries as splicing is completed, thus signifying those regions of the RNA that have been properly spliced.

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 per cent 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 most of the RNA in cells performs structural and catalytic functions (see Table 6–1, p. 306). 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 all RNAs in the cell are synthesized by a single RNA polymerase—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 discussed 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. Adequate quantities of ribosomal RNAs can be produced only because the cell 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

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 was shown in Figure 6–9. (From V.E. Foe, *Cold Spring Harbor Symp. Quant. Biol.* 42:723–740, 1978.)

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 RNA) is synthesized from a separate cluster of genes by a different polymerase, RNA polymerase III, and does not require chemical modification. It is not known why this one RNA is transcribed separately.

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 6–43A). The functions of these modifications are not understood in detail, but they probably aid in the folding and assembly of the final rRNAs and may also subtly alter the function of ribosomes. Each modification is made at a specific position in the precursor rRNA. These positions are specified by several hundred "guide RNAs," which locate 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. 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,

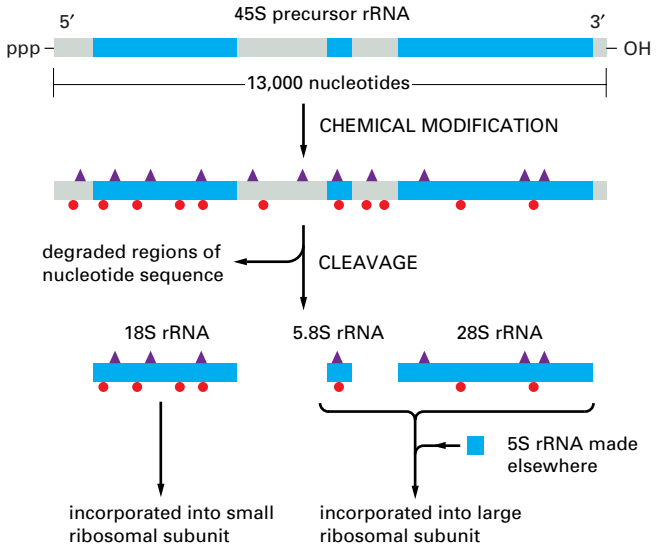


Figure 6–42 The chemical modification and nucleolytic processing of a eucaryotic 45S precursor rRNA molecule into three separate ribosomal RNAs. As indicated, two types of chemical modifications (shown in Figure 6–43) are made to the precursor rRNA before it is cleaved. Nearly half of the nucleotide 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 ultra-centrifuge. The larger the S value, the larger the rRNA.

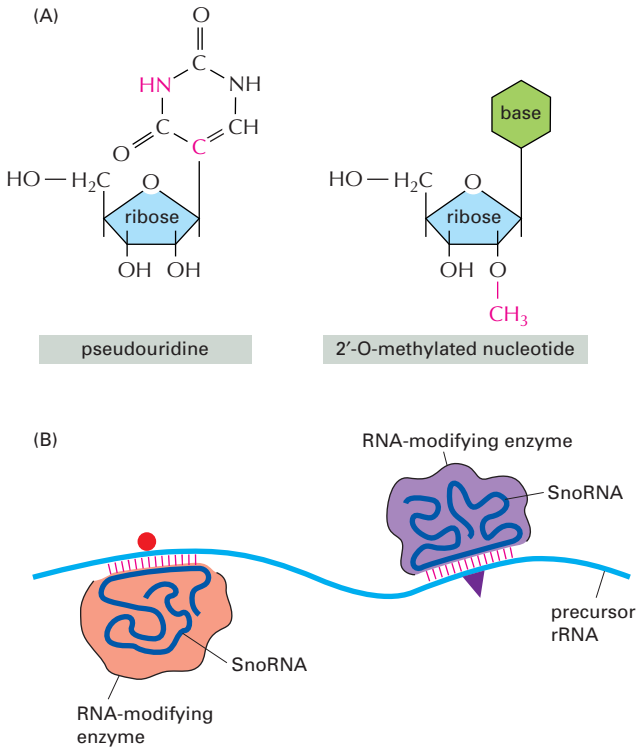


Figure 6–43 Modifications of the precursor rRNA by guide RNAs. (A) Two prominent covalent modifications occur after rRNA synthesis; the differences from the initially incorporated nucleotide are indicated by *red* atoms. (B) As indicated, snoRNAs locate the sites of modification by base-pairing to complementary sequences on the precursor rRNA. The snoRNAs are bound to proteins, and the complexes are called snoRNPs. snoRNPs contain the RNA modification activities, presumably contributed by the proteins but possibly by the snoRNAs themselves.

especially those encoding ribosomal proteins. They are therefore synthesized by RNA polymerase II and processed from excised intron sequences.

The Nucleolus Is a Ribosome-Producing Factory

The nucleolus is the most obvious structure seen in the nucleus of a 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 ribosomes. Unlike other organelles in the cell, it 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 protein subunits and partly assembled

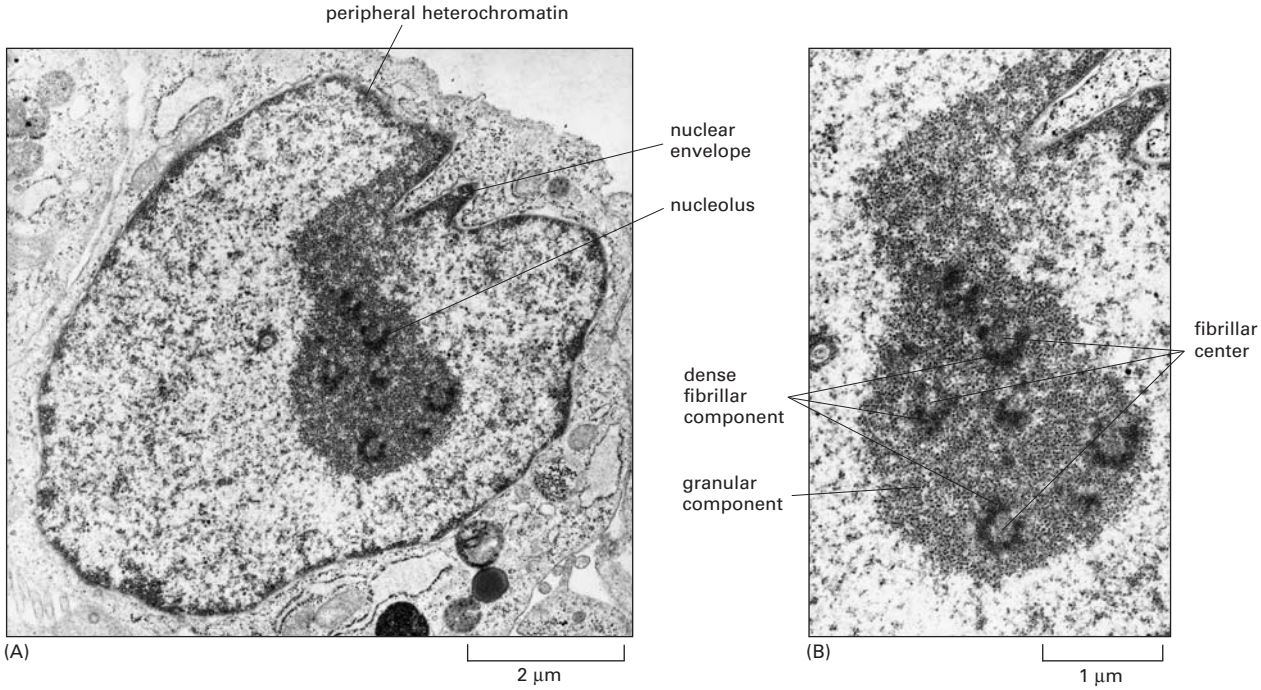
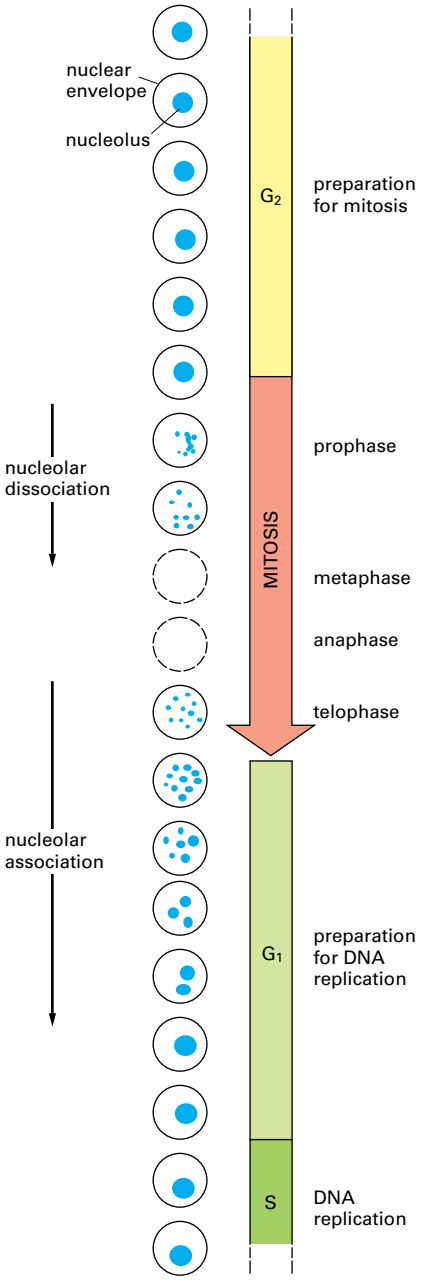


Figure 6–44 Electron micrograph of a thin section of a nucleolus in a human fibroblast, showing its three distinct zones. (A) View of entire nucleus. (B) High-power view of the nucleolus. It is believed that transcription of the rRNA genes takes place between the fibrillar center and the dense fibrillar component and that processing of the rRNAs and their assembly into ribosomes 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 membrane breaks down during mitosis, as indicated by the dashed circles.



ribosomes. The close association of all these components presumably allows the assembly of ribosomes to occur rapidly and smoothly.

It is not yet understood how the nucleolus is held together and organized, but various types of RNA molecules play a central part in its chemistry and structure, suggesting that the nucleolus 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, each of which is located near the tip of one of the two copies of five different chromosomes (see Figure 4–11). Each time a human cell undergoes mitosis, the chromosomes disperse and the nucleolus breaks up; after mitosis, the tips of the 10 chromosomes coalesce as the nucleolus reforms (Figures 6–45 and 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.

A schematic diagram of the assembly of ribosomes is shown 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, as we have seen, 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. Thus, the nucleolus can be thought of as a large factory at which many different noncoding RNAs are 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 visualized and studied (Figure 6–48). These include Cajal bodies (named for the scientist who first described them in 1906), GEMS (Gemini of coiled bodies), and interchromatin granule clusters (also called "speckles"). Like the nucleolus, these other nuclear structures lack membranes

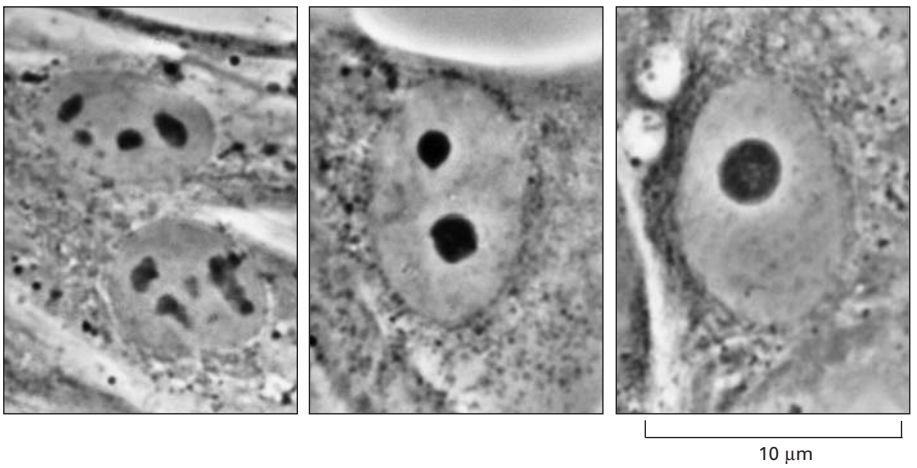


Figure 6–46 Nucleolar fusion. These light micrographs of human fibroblasts grown in culture show various stages of nucleolar fusion. After mitosis, each of the ten human chromosomes that carry a cluster of rRNA genes begins to form a tiny nucleolus, but these rapidly coalesce as they grow to form the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

and are highly dynamic; their appearance is probably the result of the tight association of protein and RNA (and perhaps DNA) 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 are truly distinct structures. They may be sites where snRNAs and snoRNAs undergo their final modifications and assembly with protein. Both the RNAs and the proteins that make up the snRNPs are partly assembled in the cytoplasm, but they are transported into the nucleus for their final modifications. It has been proposed that Cajal bodies/GEMS are also sites where the snRNPs are recycled and their RNAs are "reset" after the rearrangements that occur during splicing (see p. 322). In contrast, the interchromatin granule clusters have been proposed to be stockpiles of fully mature snRNPs that are ready to be used in splicing of pre-mRNAs (Figure 6–49).

Scientists have had difficulties in working out the function of the small subnuclear structures just described. Much of the progress now being made depends on genetic tools—examination of the effects of designed mutations in mice 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

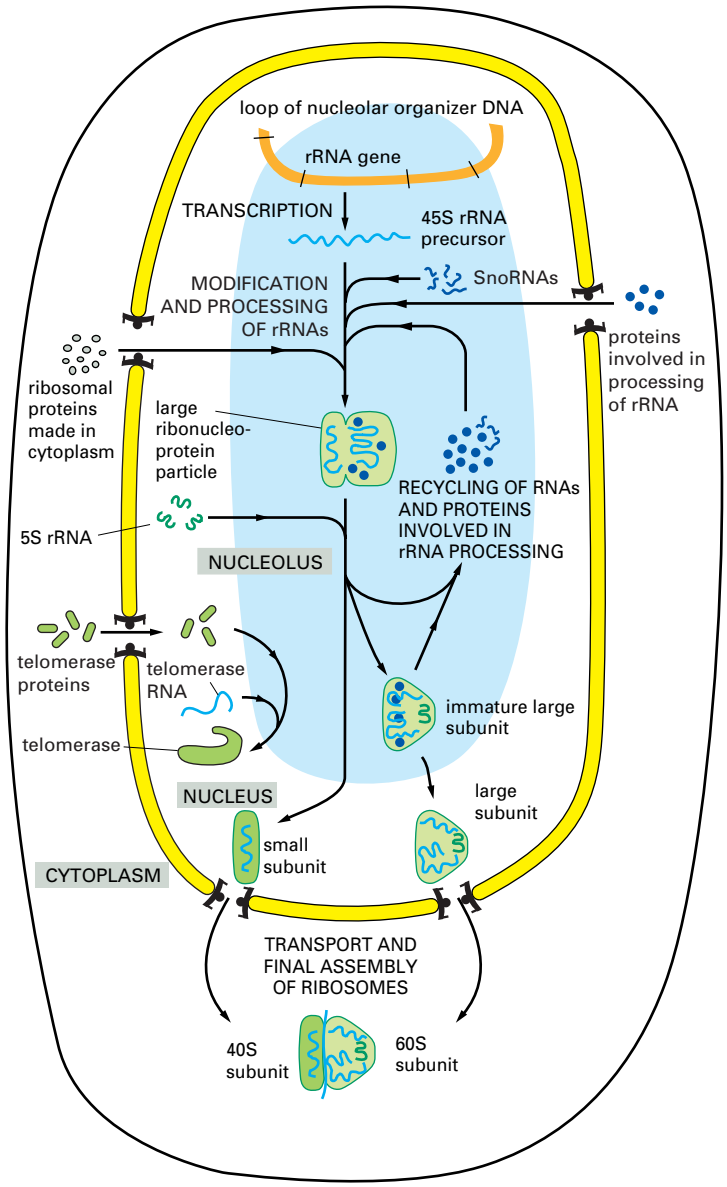


Figure 6–47 The function of the nucleolus in ribosome and other ribonucleoprotein synthesis. The 45S precursor rRNA is packaged in a large ribonucleoprotein particle containing many ribosomal proteins imported from the cytoplasm. While this particle remains in the nucleolus, selected pieces are added 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.

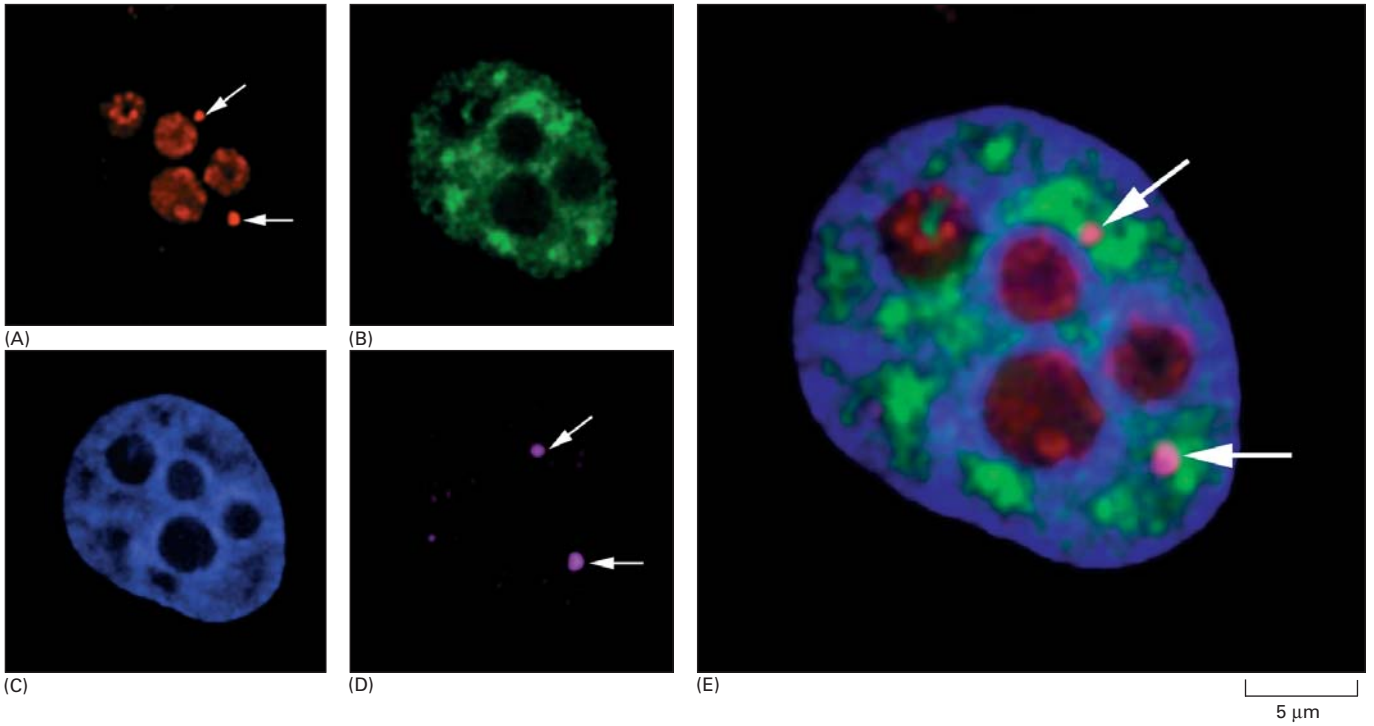


Figure 6–48 Visualization of chromatin and nuclear bodies. (A)–(D) show micrographs of the same human cell nucleus, each processed differently to show a particular set of nuclear structures. (E) shows an enlarged superposition of all four individual images. (A) shows the location of the protein fibrillarin (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 (indicated by arrows). (From J.R. Swedlow and A.I. Lamond, *Gen. Biol.* 2:1–7, 2001; micrographs courtesy of Judith Sleeman.)

to be caused by a subtle defect in snRNP assembly and subsequent pre-mRNA splicing. More severe defects would be expected to be lethal.

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,

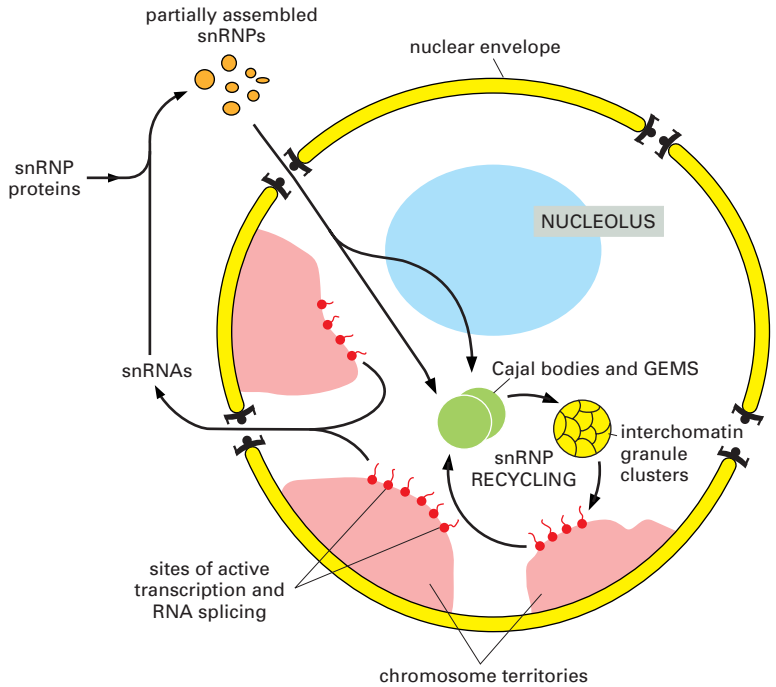


Figure 6–49 Schematic view of subnuclear structures. A typical vertebrate nucleus has several Cajal bodies, which are proposed to be the sites where snRNPs and snoRNPs undergo their final modifications. Interchromatin granule clusters are proposed to be storage sites for fully mature snRNPs. A typical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

After their initial synthesis, snRNAs are exported from the nucleus, after which they 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 in Cajal bodies. In addition, the U6 snRNP requires chemical modification by snoRNAs in 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.)

we have seen that the assembly of splicing components on pre-mRNA is co-transcriptional; thus splicing must occur at many locations along chromosomes. We saw in Chapter 4 that interphase chromosomes occupy discrete territories in the nucleus, and transcription and pre-mRNA splicing must take place within these territories. However, interphase chromosomes are themselves dynamic and their exact positioning in the nucleus correlates with gene expression. For example, transcriptionally silent regions of interphase chromosomes are often associated with the nuclear envelope where the concentration of heterochromatin components is believed to be especially high. When these same regions become transcriptionally active, they relocate towards the interior of the nucleus, which is richer in the components required for mRNA synthesis. It has been proposed that, 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" where the local concentration of these components is very high. As a result, 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 (Figure 6–49).

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—designated polymerase I, II, and III—that are related evolutionarily to one another and to the bacterial polymerase.

Eucaryotic mRNA is synthesized by RNA polymerase II. This enzyme requires a series of additional proteins, termed the general transcription factors, to initiate transcription on a purified DNA template and still more proteins (including chromatin-remodeling complexes and histone acetyltransferases) to initiate transcription on its chromatin template 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). Some of these RNA processing events that modify the initial RNA transcript (for example, those involved in RNA splicing) are carried out primarily by special small RNA molecules.

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 ribosomes 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 first attracted the attention of biologists in the late 1950s,

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

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 subunits—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 elaborate machinery by which cells read this code—the ribosome—was finally revealed in atomic detail.

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. 132–133, 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.

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 four different nucleotides in mRNA and twenty 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 medium 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 consecutively in 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).

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.

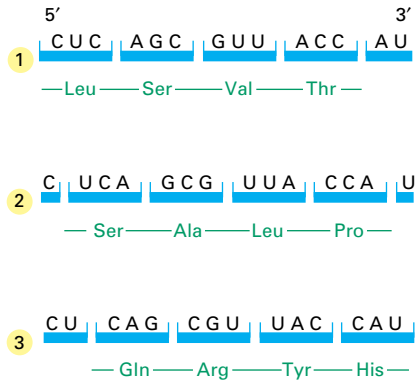


Figure 6–51 The three possible reading frames in protein synthesis. In the process of translating a nucleotide sequence *(blue)* into an amino acid sequence *(green)*, the sequence of nucleotides in an mRNA molecule is read from the 5′ to the 3′ end in sequential sets of three nucleotides. In principle, 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

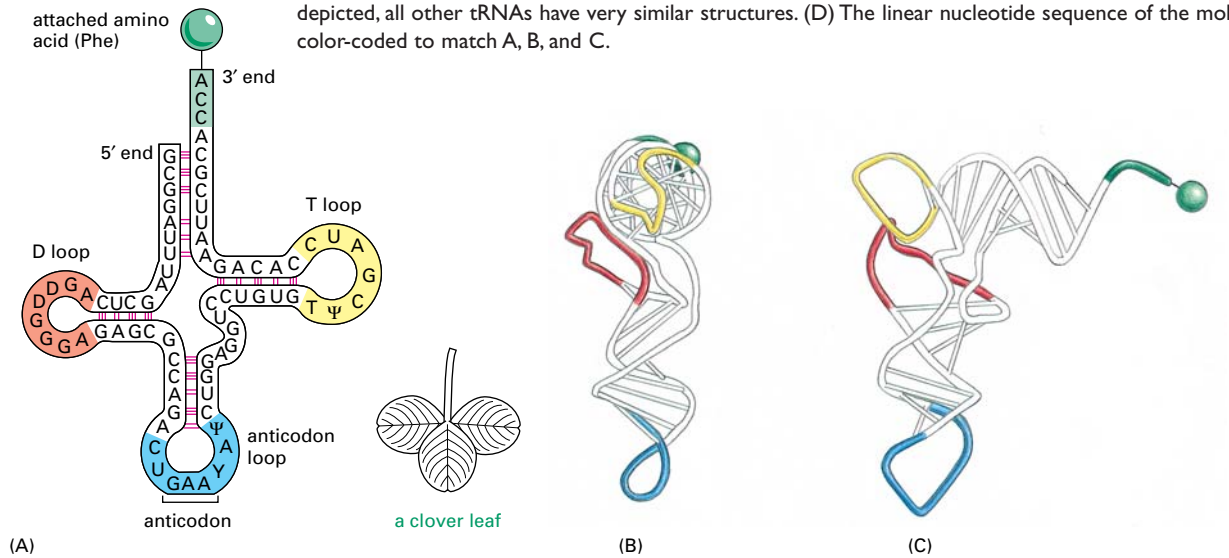
The codons in an mRNA molecule do not directly recognize the amino acids they specify: the group of three nucleotides does not, for example, bind directly to the amino acid. Rather, the translation of mRNA into protein depends on adaptor molecules that can recognize and bind both to the codon and, at 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 up into precisely defined 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 6–52A). 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 (Figure 6–52B,C).

Two regions of unpaired nucleotides situated at either end of the L-shaped molecule are crucial to the function of tRNA in protein synthesis. One of these regions forms the **anticodon**, a set of three consecutive nucleotides that pairs with the complementary codon in an mRNA molecule. The other is a short single-stranded region at the 3′ end of the molecule; this is the site where the amino acid that matches the codon is attached to the tRNA.

We 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

Figure 6–52 A tRNA molecule. In this series of diagrams, the same tRNA molecule—in this case a tRNA specific for the amino acid phenylalanine (Phe)—is depicted in various ways. (A) The cloverleaf structure, a convention used to show 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 (for pseudouridine—see Figure 6–43) and D (for dihydrouridine—see Figure 6–55) are derived from uracil. (B and C) Views of the actual L-shaped molecule, based on x-ray diffraction analysis. Although a particular tRNA, that for the amino acid phenylalanine, is depicted, all other tRNAs have very similar structures. (D) The linear nucleotide sequence of the molecule, color-coded to match A, B, and C.



5′ GCGGAUUUAGCUCAGDDGGGAGAGCGCCAGACUGAAYAΨCUGGAGGUCCUGUGTΨCGAUCCACAGAAUUCGCACCA 3′

(D) anticodon

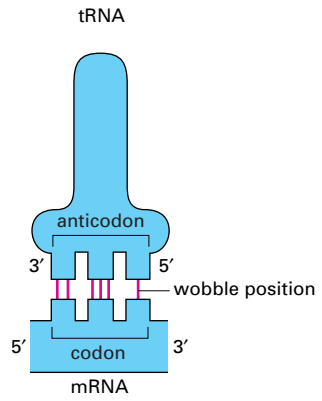


Figure 6–53 Wobble base-pairing between codons and anticodons. If the nucleotide listed in the first column is present at the third, or wobble, position of the codon, it can base-pair with any of the nucleotides listed in the second column. Thus, for example, when inosine (I) is present in the wobble 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 which 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.)

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	G or I
C	G or I
A	U
G	C

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 497 tRNA genes but, among them, only 48 different anticodons are represented.

tRNAs Are Covalently Modified Before They Exit from the Nucleus

We have seen that most eucaryotic RNAs are covalently altered before they are allowed to exit from the nucleus, and tRNAs are no exception. Eucaryotic tRNAs are synthesized by RNA polymerase III. Both bacterial and eucaryotic tRNAs are typically synthesized as larger precursor tRNAs, and these 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 is chemically distinct from that of pre-mRNA splicing; rather than generating a lariat intermediate, tRNA splicing occurs through 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

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.)

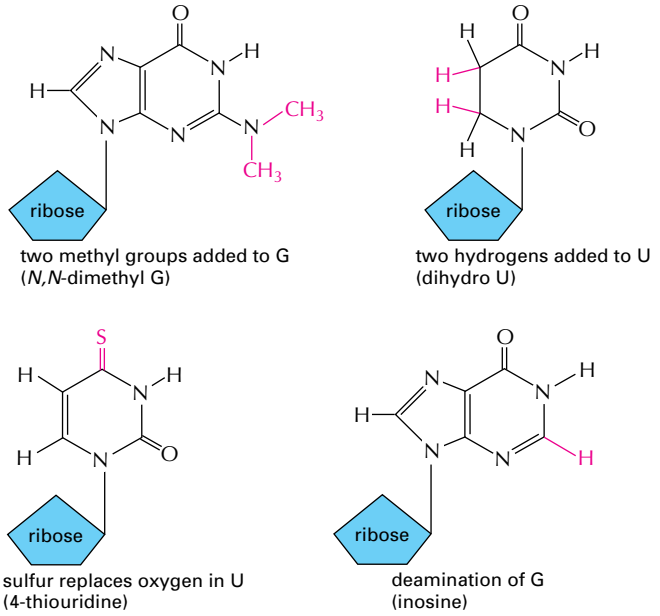


Figure 6–55 A few of the unusual nucleotides found in tRNA molecules. These nucleotides are produced by covalent modification of a normal nucleotide after it has been incorporated into an RNA chain. In most tRNA molecules about 10% of the nucleotides are modified (see Figure 6–52).

trimming and splicing reactions are thought to act as quality-control steps in the generation of tRNAs.

All tRNAs are also subject to a variety of chemical modifications—nearly one 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 guanosine—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.

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 (Figures 6–56 and 6–57). For most cells there is a different synthetase enzyme for each amino acid (that is, 20 synthetases in all); one attaches glycine

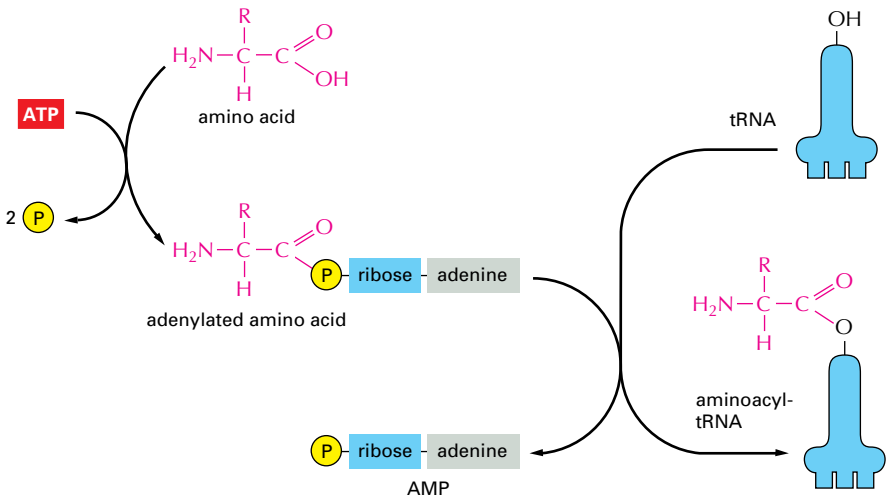


Figure 6–56 Amino acid activation. The two-step process in which an amino acid (with its side chain denoted by R) is activated for protein synthesis by an aminoacyl-tRNA synthetase enzyme is shown. As indicated, the energy of ATP hydrolysis is used to attach each amino acid to its tRNA molecule in a 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.

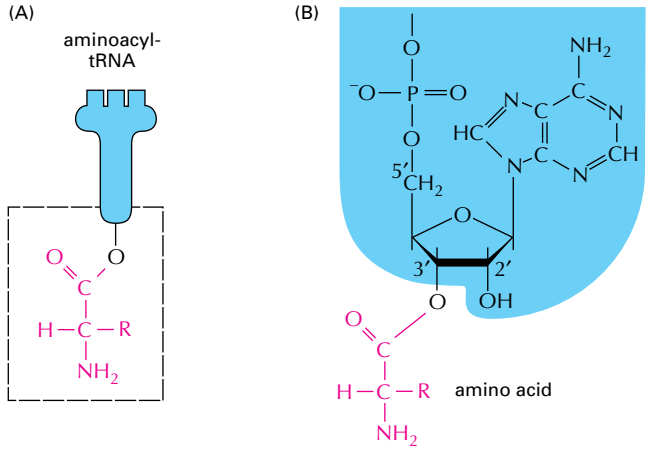


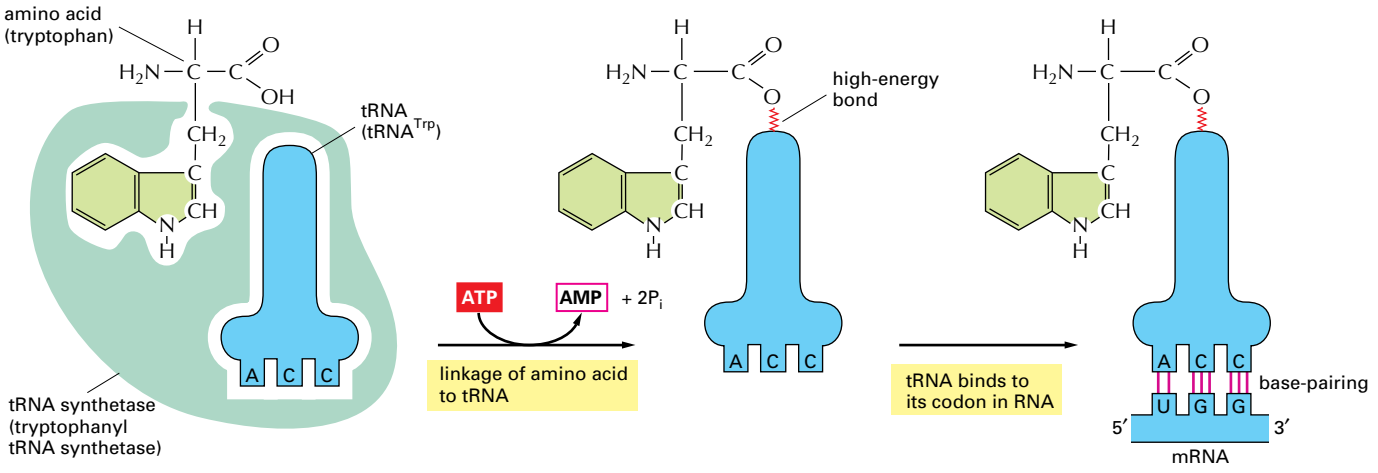
Figure 6–57 The structure of the aminoacyl-tRNA linkage. The carboxyl end of the amino acid forms an ester bond to ribose. Because the hydrolysis of this ester bond is associated with a large favorable change in free energy, an amino acid held in this way is said to be activated. (A) Schematic drawing 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 boxed region in (A). These 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.

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.

The synthetase-catalyzed reaction that attaches the amino acid to the 3′ end of the tRNA is one of many cellular reactions coupled to the energy-releasing hydrolysis of ATP (see pp. 83–84), 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.

Although the tRNA molecules serve as the final adaptors in converting nucleotide sequences into amino acid sequences, the aminoacyl-tRNA synthetase enzymes are adaptors of equal importance in the decoding process (Figure 6–58). This was established by an ingenious experiment in which an 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 cells have several quality control mechanisms to avoid this type of mishap, the experiment clearly 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

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.



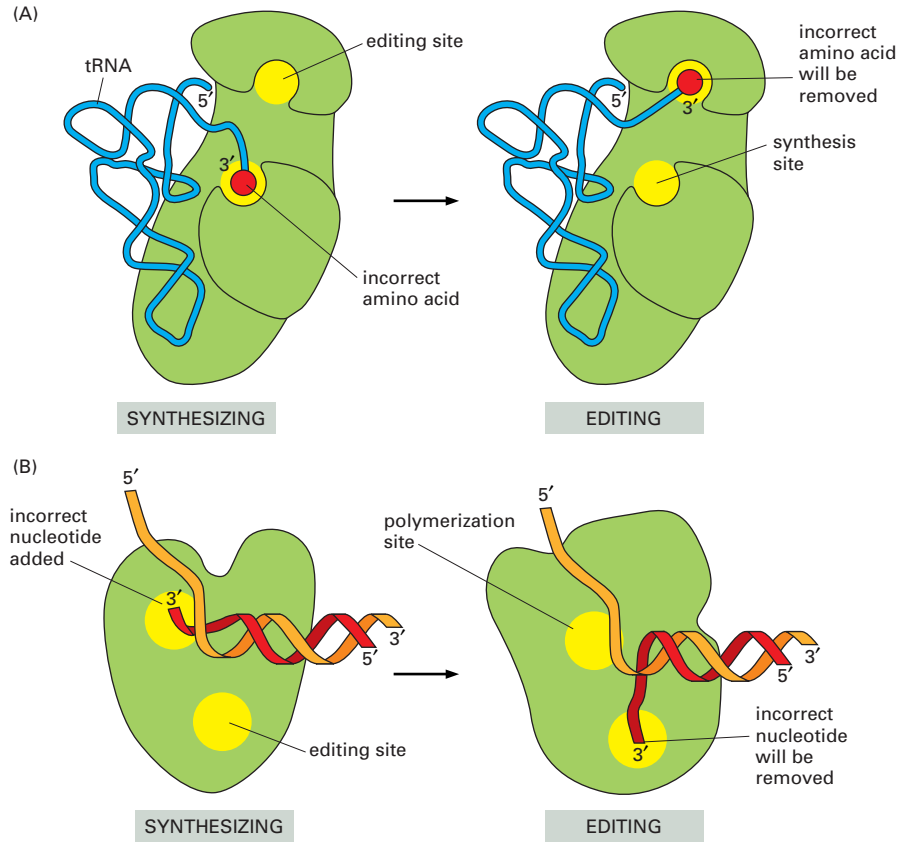


Figure 6–59 Hydrolytic editing. (A) tRNA synthetases remove their own coupling errors through hydrolytic editing of incorrectly attached amino acids. As described in the text, the correct amino acid is rejected by the editing site. (B) The error-correction process performed by DNA polymerase shows some similarities; however, it differs so far as the removal process depends strongly on a mispairing with the template (see Figure 5–9).

action that associates each sequence of three nucleotides in the mRNA molecule—that is, each codon—with its particular amino acid.

Editing by RNA 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 do so by a two-step mechanism. First, the correct amino acid has the highest affinity for the active-site pocket of its synthetase and is therefore favored over the other 19. In particular, amino acids larger than the correct one are 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 forces 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 released from the enzyme. This hydrolytic editing, which is analogous to the editing by DNA polymerases (Figure 6–59), raises the overall accuracy of tRNA charging to approximately one mistake in 40,000 couplings.

The tRNA synthetase must also recognize the correct set of tRNAs, and extensive structural and chemical complementarity between the synthetase and the tRNA allows various features of the tRNA to be sensed (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 the nucleotide in the anticodon. For other synthetases it is the nucleotide sequence of the acceptor stem that is the key recognition determinant. In most cases, however, nucleotides at several positions on the tRNA are "read" by the synthetase.

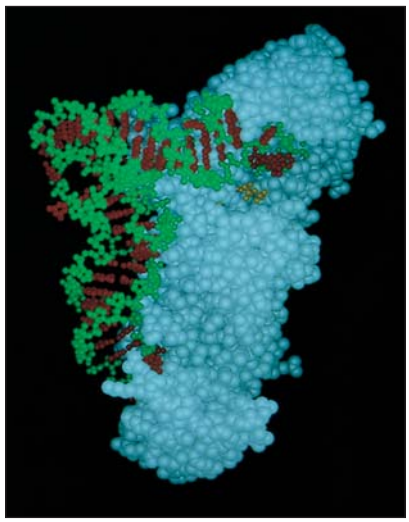


Figure 6–60 The recognition of a tRNA molecule by its aminoacyl-tRNA synthetase. For this tRNA ($tRNA^{Gln}$), specific nucleotides in both the anticodon *(bottom)* and the amino acid-accepting arm allow the correct tRNA to be recognized by the synthetase enzyme *(blue)*. (Courtesy of Tom Steitz.)

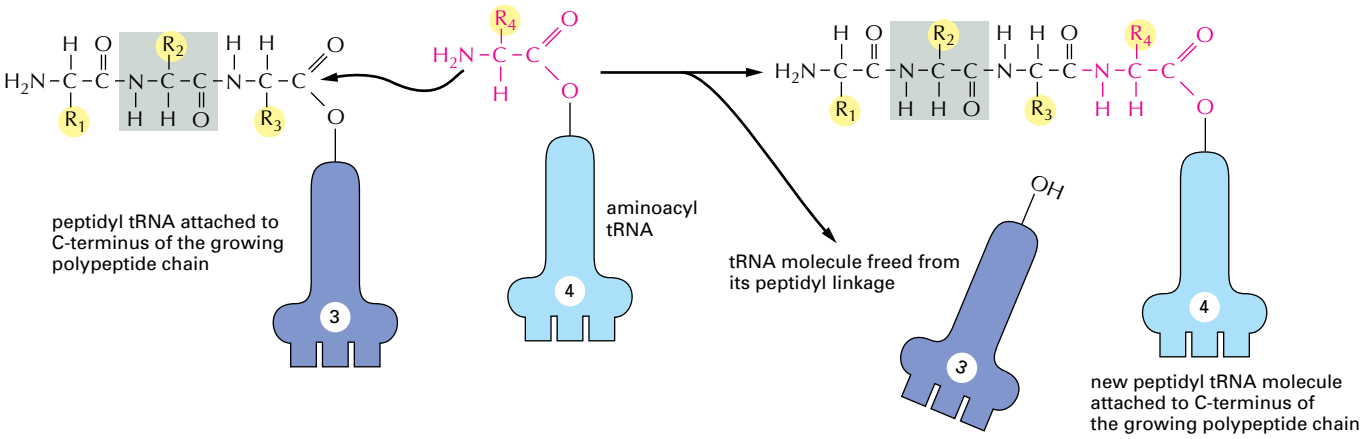


Figure 6–61 The incorporation of an amino acid into a protein. A polypeptide chain grows by the stepwise addition of amino acids to its C-terminal end. The formation of each peptide bond is energetically favorable because the growing C-terminus has been activated by the covalent attachment of a tRNA 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 to the growing chain.

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 by which they are joined 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 (a peptidyl-tRNA molecule). This high-energy covalent linkage is disrupted during each addition but is immediately replaced by the 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 on Ribosomes

As we have seen, the synthesis of proteins is guided by information carried by mRNA molecules. To maintain the correct reading frame and to ensure accuracy (about 1 mistake every 10,000 amino acids), protein synthesis is performed in the **ribosome**, a complex catalytic machine made from more than 50 different proteins (the *ribosomal proteins*) and several RNA molecules, the **ribosomal RNAs (rRNAs)**. A typical eucaryotic cell contains millions of ribosomes in its cytoplasm (Figure 6–62). As we have seen, eucaryotic ribosomal subunits are

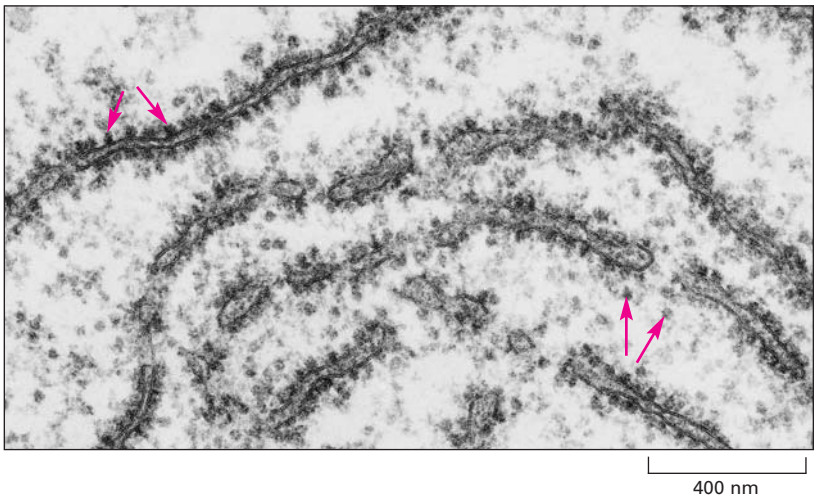


Figure 6–62 Ribosomes in the cytoplasm of a eucaryotic cell. This electron micrograph shows a thin section of a small region of cytoplasm. The ribosomes appear as black dots *(red arrows)*. Some are free in the cytosol; others are attached to membranes of the endoplasmic reticulum. (Courtesy of Daniel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

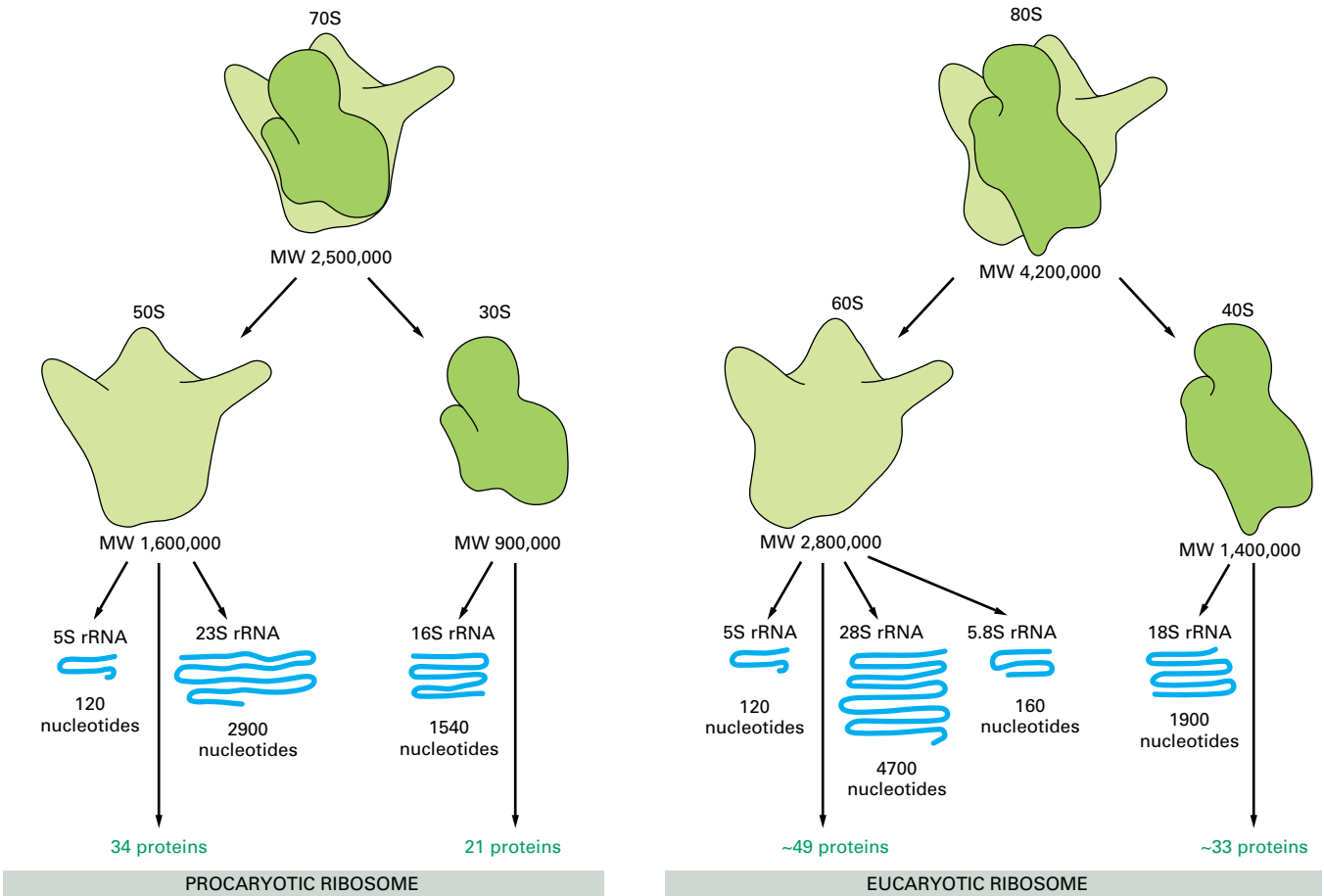
assembled at the nucleolus, by the association of newly transcribed and modified rRNAs with ribosomal proteins, which have been transported into the nucleus after their synthesis in the cytoplasm. The two ribosomal subunits are then exported to the cytoplasm, where they perform protein synthesis.

Eucaryotic and procaryotic ribosomes are very similar in design and function. 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 a framework on which the tRNAs can be accurately 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 encounter the ribosome's active site, 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, 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

Figure 6–63 A comparison of the structures of procaryotic and eucaryotic ribosomes. Ribosomal components are commonly designated by their "S values," which refer to their rate of sedimentation in an ultracentrifuge. Despite the 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 extra 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.



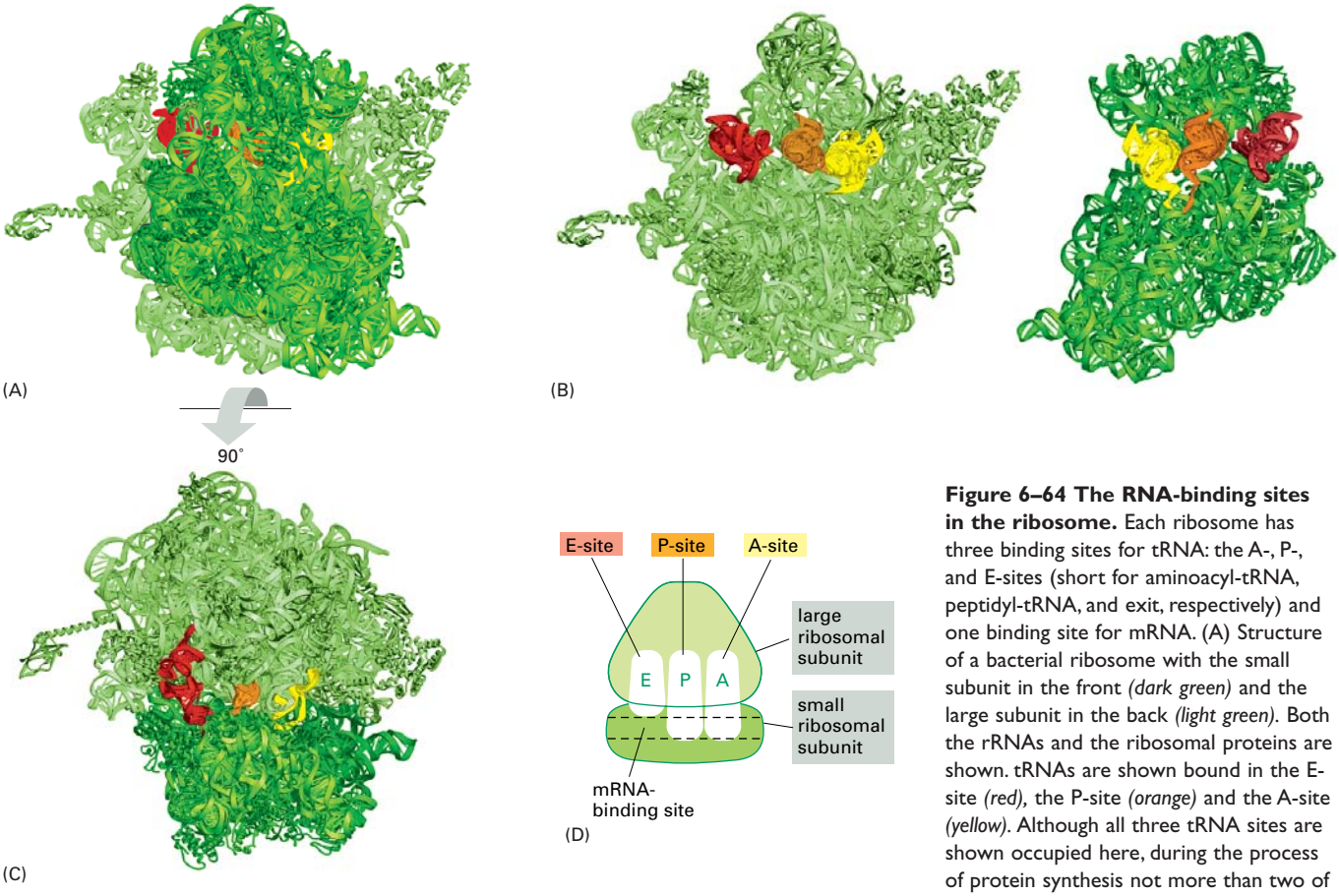


Figure 6–64 The RNA-binding sites in the ribosome. Each ribosome has three binding sites for tRNA: the A-, P-, and E-sites (short for aminoacyl-tRNA, peptidyl-tRNA, and exit, respectively) and one binding site for mRNA. (A) Structure of a bacterial ribosome with the small subunit in the front *(dark 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 6–65). (B) Structure of the large *(left)* and small *(right)* ribosomal subunits arranged as though the ribosome in (A) were opened like a book. (C) Structure of the ribosome in (A) viewed from the top. (D) Highly 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, courtesy of Albion Bausom and Harry Noller.)

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 bound to the ribosome. 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 three major steps. 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 (Figure 6–65). 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 codon in mRNA 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* catalytic activity contained in the large ribosomal subunit. This reaction is accompanied by several conformational changes in the ribosome, which shift the two tRNAs into the E- and P-sites of the large subunit. In step 3, another series of conformational changes moves the mRNA exactly three nucleotides through the ribosome and resets the ribosome so it is ready to receive the next amino acyl tRNA. Step 1 is then repeated with a new incoming aminoacyl tRNA, and so on.

This three-step cycle is repeated each time an amino acid is added to the polypeptide chain, and the chain grows from its amino to its carboxyl end until a stop codon is encountered.

Elongation Factors Drive Translation Forward

The basic cycle of polypeptide elongation shown in outline in Figure 6–65 has an additional feature that makes translation especially efficient and accurate. Two *elongation factors* (EF-Tu and EF-G) enter and leave the ribosome during each cycle, each hydrolyzing GTP to GDP and undergoing conformational changes in the process. Under some conditions, ribosomes can be made to perform protein synthesis without the aid of the elongation factors and GTP hydrolysis, but this synthesis is very slow, inefficient, and inaccurate. The process is speeded up enormously by coupling conformational changes in the elongation factors to transitions between different conformational states of the ribosome. Although these conformational changes in the ribosome are not yet understood in detail, some may involve RNA rearrangements similar to those occurring in the RNAs of the spliceosome (see Figure 6–30). The cycles of elongation factor association, GTP hydrolysis, and dissociation ensures that the conformational changes occur in the "forward" direction and translation thereby proceeds efficiently (Figure 6–66).

In addition to helping move translation forward, EF-Tu is thought to increase the accuracy of translation by monitoring the initial interaction between a charged tRNA and a codon (see Figure 6–66). Charged tRNAs enter the ribosome bound to the GTP-form of EF-Tu. Although the bound elongation factor allows codon–anticodon pairing to occur, it prevents the amino acid from being incorporated into the growing polypeptide chain. The initial codon recognition, however, triggers the elongation factor to hydrolyze its bound GTP (to GDP and inorganic phosphate), whereupon the factor dissociates from the ribosome without its tRNA, allowing protein synthesis to proceed. The elongation factor introduces two short delays between codon–anticodon base pairing and polypeptide chain elongation; these delays selectively permit incorrectly bound tRNAs to exit from the ribosome before the irreversible step of chain elongation occurs. The first delay is the time required for GTP hydrolysis. The rate of GTP hydrolysis by EF-Tu is faster for a correct codon–anticodon pair than for an incorrect pair; hence an incorrectly bound tRNA molecule has a longer window of opportunity to dissociate from the ribosome. In other words, GTP hydrolysis selectively captures the correctly bound tRNAs. A second lag occurs between EF-Tu dissociation and the full accommodation of the tRNA in the A site of the ribosome. Although this lag is believed to be the same for correctly and incorrectly bound tRNAs, an incorrect tRNA molecule forms a smaller number of codon–anticodon hydrogen bonds than does a correctly matched pair and is therefore more likely to dissociate during this period. These two delays introduced by the elongation factor cause most incorrectly bound tRNA molecules (as well as a significant number of correctly bound molecules) to leave the

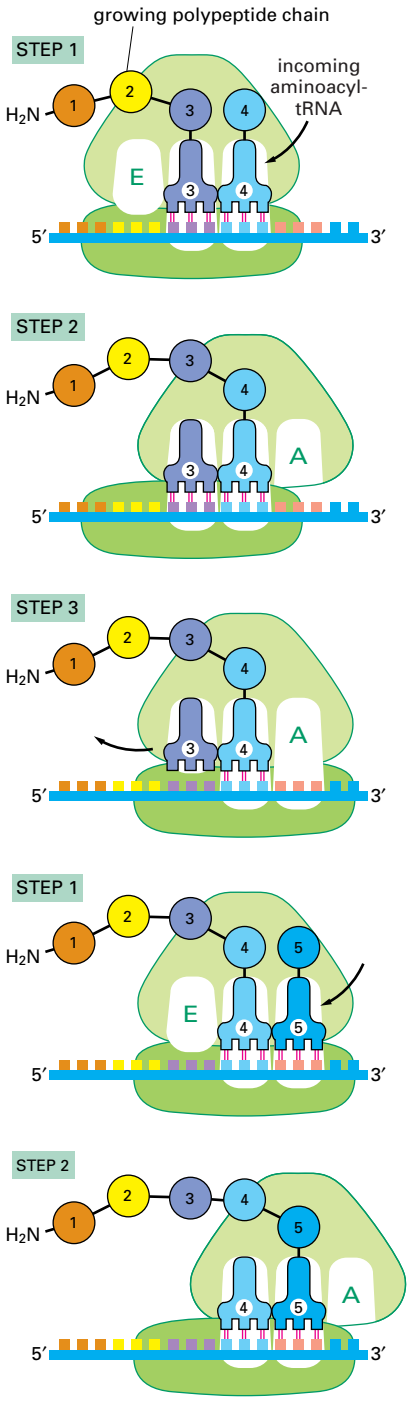
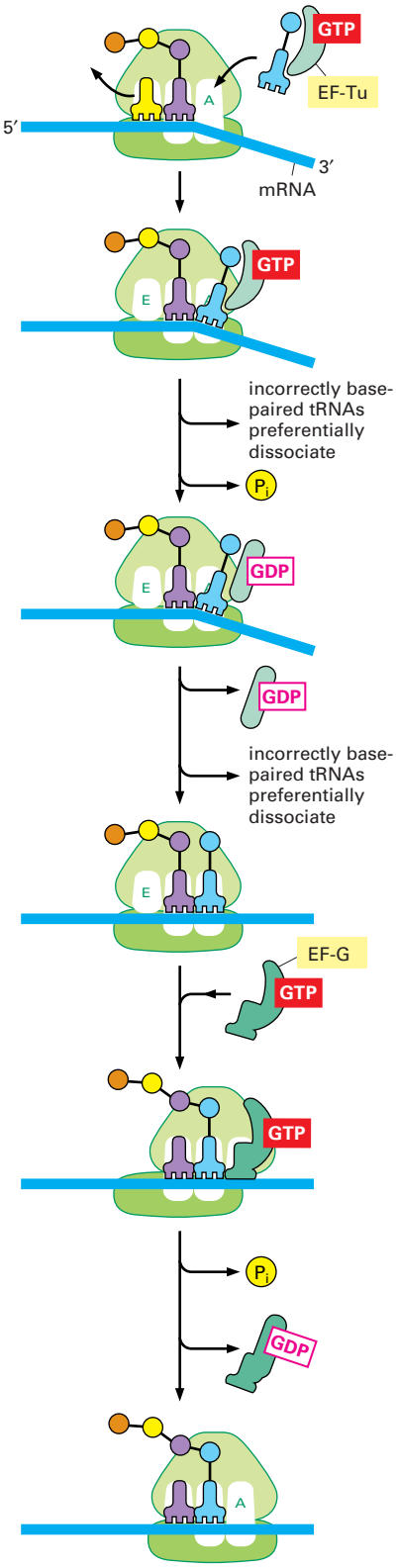


Figure 6–65 Translating an mRNA molecule. Each amino acid added to the growing end of a polypeptide chain is selected by complementary base-pairing between the anticodon on its attached tRNA molecule and the next codon on the mRNA chain. Because only one of the many types of tRNA molecules in a cell can base-pair with each codon, the codon determines the specific amino acid to be added to the growing polypeptide chain. The three-step cycle shown is repeated over and over during the synthesis of a protein. An aminoacyl-tRNA molecule binds to a vacant A-site on the ribosome in step 1, a new peptide bond is formed in step 2, and the mRNA moves a distance of three nucleotides through the small-subunit chain in step 3, ejecting the spent tRNA molecule and "resetting" the ribosome so that the next aminoacyl-tRNA molecule can bind. Although the figure shows a large movement of the small ribosome subunit relative to the large subunit, the conformational changes that actually take place in the ribosome during translation are more subtle. It is likely that they involve a series of small rearrangements within each subunit as well as several small shifts between the two subunits. 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. The position at which the growing peptide chain is attached to a tRNA does not change during the elongation cycle: it is always linked to the tRNA present in the P site of the large subunit.

Figure 6–66 Detailed view of the translation cycle. The outline of translation presented in Figure 6–65 has been supplemented with additional features, including the participation of elongation factors and a mechanism by which translational accuracy is improved. In the initial binding event *(top panel)* an aminoacyl-tRNA molecule that is tightly bound to EF-Tu pairs transiently with the codon at the A-site in the small subunit. During this step *(second panel),* the tRNA occupies a hybrid-binding site on the ribosome. The codon–anticodon pairing triggers GTP hydrolysis by EF-Tu causing it to dissociate from the aminoacyl-tRNA, which now enters the A-site *(fourth panel)* and can participate in chain elongation. A delay between aminoacyl-tRNA binding and its availability for protein synthesis is thereby inserted into the protein synthesis mechanism. As described in the text, this delay increases the accuracy of translation. In subsequent steps, elongation factor EF-G in the GTP-bound form enters the ribosome and binds in or near the A-site on the large ribosomal subunit, accelerating the movement of the two bound tRNAs into the A/P and P/E hybrid states. Contact with the ribosome stimulates the GTPase activity of EF-G, causing a dramatic conformational change in EF-G as it switches from the GTP to the GDP-bound form. This change moves the tRNA bound to the A/P hybrid state to the P-site and advances the cycle of translation forward by one codon.

During each cycle of translation elongation, the tRNAs molecules move through the ribosome in an elaborate series of gyrations during which they transiently occupy several "hybrid" binding states. In one, the tRNA is simultaneously bound to the A site of the small subunit and the P site of the large subunit; in another, the tRNA is bound to the P site of the small subunit and the E site of the large subunit. In a single cycle, a tRNA molecule is considered to occupy six different sites, the initial binding site (called the A/T hybrid state), the A/A site, the A/P hybrid state, the P/P site, the P/E hybrid state, and the E-site. Each tRNA is thought to ratchet through these positions, undergoing rotations along its long axis at each change in location.

EF-Tu and EF-G are the designations used for the bacterial elongation factors; in eucaryotes, they are called EF-1 and EF-2, respectively. The dramatic change in the three-dimensional structure of EF-Tu that is caused by GTP hydrolysis was illustrated in Figure 3–74. For each peptide bond formed, a molecule of EF-Tu and EF-G are each released in their inactive, GDP-bound forms. To be used again, these proteins must have their GDP exchanged for GTP. In the case of EF-Tu, this exchange is performed by a specific member of a large class of proteins known as *GTP exchange factors*.



ribosome without being used for protein synthesis, and this two-step mechanism is largely responsible for the 99.99% accuracy of the ribosome in translating proteins.

Recent discoveries indicate that EF-Tu may have an additional role in raising the overall accuracy of translation. Earlier in this chapter, we discussed the key role of aminoacyl synthetases in accurately matching amino acids to tRNAs. As the GTP-bound form of EF-Tu escorts aminoacyl-tRNAs to the ribosome (see Figure 6–66), it apparently double-checks for the proper correspondence between amino acid and tRNA and rejects those that are mismatched. Exactly how this is accomplished is not well-understood, but it may involve the overall binding energy between EF-Tu and the aminoacyl-tRNA. According to this idea, correct matches have a narrowly defined affinity for EF-Tu, and incorrect matches bind either too strongly or too weakly. EF-Tu thus appears to discriminate, albeit crudely, among many different amino acid-tRNA combinations, selectively allowing only the correct ones to enter the ribosome.

The Ribosome Is a Ribozyme

The ribosome is a very large and complex structure, composed of two-thirds RNA and one-third protein. The determination, in 2000, of the entire three-dimensional structure of its large and small subunits is a major triumph of modern structural biology. The structure strongly confirms the 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. Thus, for example, the ribosomal RNAs are folded into

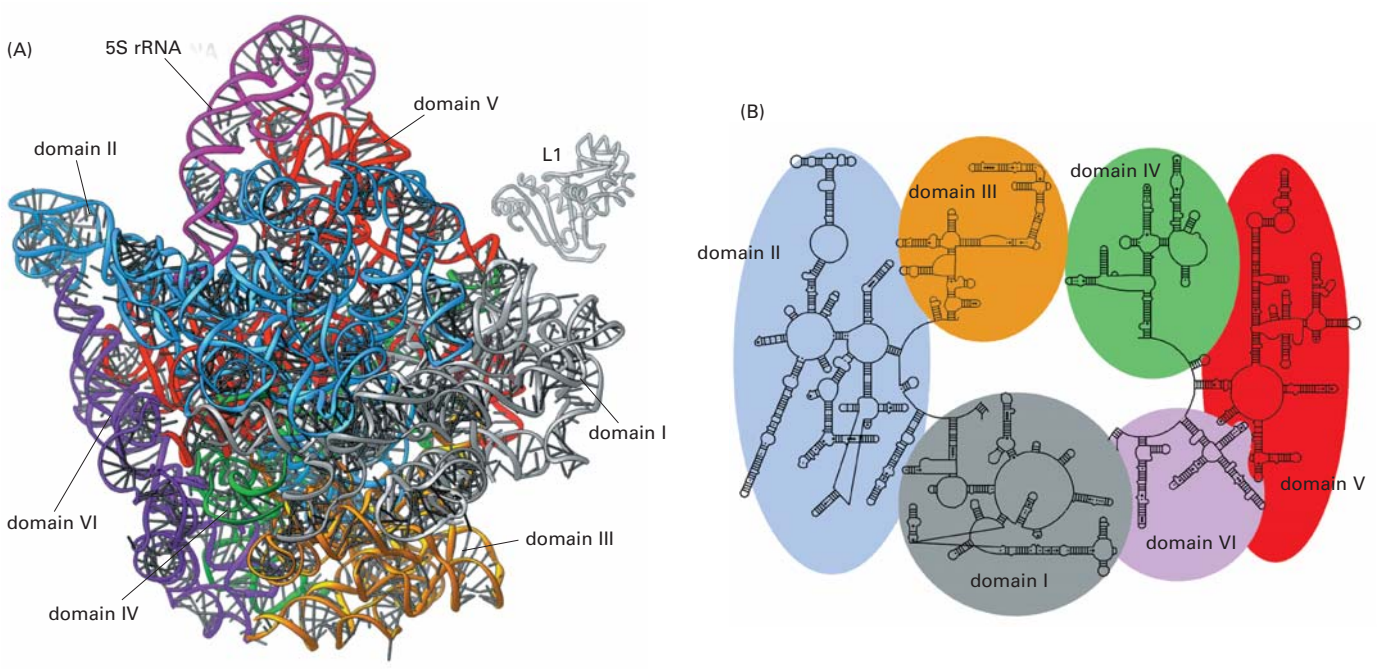


Figure 6–67 Structure of the rRNAs in the large subunit of a bacterial ribosome, as determined by x-ray crystallography. (A) Three-dimensional structures of the large-subunit rRNAs (5S and 23S) as they appear in the ribosome. One of the protein subunits of the ribosome (L1) is also shown as a reference 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 structural 'domains' whose colors correspond to those of the three-dimensional structure 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.)

highly compact, precise three-dimensional structures that form the compact core of the ribosome and thereby determine its overall shape (Figure 6–67).

In marked contrast to the central positions of the rRNA, the ribosomal proteins are generally located on the surface and fill in the gaps and crevices of the folded RNA (Figure 6–68). Some of these proteins contain globular domains on the ribosome surface that send out extended regions of polypeptide chain that penetrate short distances into holes in the RNA core (Figure 6–69). The main role

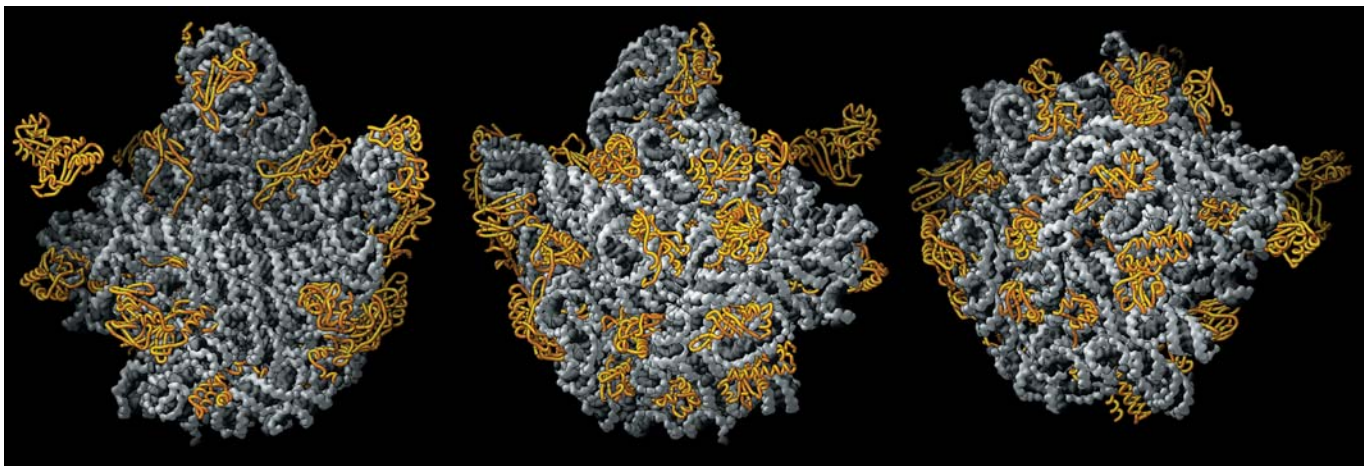


Figure 6–68 Location of the protein components of the bacterial large ribosomal subunit. The rRNAs (5S and 23S) are depicted in *gray* and the large-subunit proteins (27 of the 31 total) in *gold*. For convenience, the protein structures depict only the polypeptide backbones. (A) View of the interface with the small subunit, the same view shown in Figure 6–64B. (B) View of the back of the large subunit, obtained by rotating (A) by 180° around a vertical axis. (C) View of the bottom of the large subunit showing the peptide exit channel in the center of the structure. (From N. Ban et al., *Science* 289:905–920, 2000. © AAAS.)

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.

Not only are the three binding sites for tRNAs (the A-, P-, and E-sites) on the ribosome formed primarily by the ribosomal RNAs, but the catalytic site for peptide bond formation is clearly formed by the 23S RNA, with the nearest amino acid located more than 1.8 nm away. This RNA-based catalytic site for peptidyl transferase is similar in many respects to those found in some proteins; it is a highly structured pocket that precisely orients the two reactants (the growing peptide chain and an aminoacyl-tRNA), and it provides a functional group to act as a general acid–base catalyst—in this case apparently, a ring nitrogen of adenine, instead of an amino acid side chain such as histidine (Figure 6–70). The ability of an RNA molecule to act as such a catalyst was initially surprising because RNA was thought to lack an appropriate chemical group that could both accept and donate a proton. Although the p*K* of adenine-ring nitrogens is usually around 3.5, the three-dimensional structure and charge distribution of the 23S rRNA active site force the p*K* of this apparently critical adenine into the neutral range and thereby create the enzymatic activity.

RNA molecules that possess catalytic activity are known as **ribozymes**. We saw earlier in this chapter how other ribozymes function in RNA-splicing reactions (for example, see Figure 6–36). In the final section of this chapter, we consider what the recently recognized ability of RNA molecules to function as catalysts for a wide variety of different reactions might mean for the early evolution of living cells. Here we need only 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, might be viewed as a relic of an earlier time in life's history—when protein synthesis evolved in cells that were run almost entirely by ribozymes.

Figure 6–69 Structure of the L15 protein in the large subunit of the bacterial ribosome. The globular domain of the protein lies on the surface of the ribosome and an extended region penetrates deeply into the RNA core of the ribosome. The L15 protein is shown in *yellow* and a portion of the ribosomal RNA core is shown in *red*. (Courtesy of D. Klein, P.B. Moore and T.A. Steitz.)

Nucleotide Sequences in mRNA Signal Where to Start Protein Synthesis

The initiation and termination of translation occur through variations on 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, so that a nonfunctional protein with a garbled sequence of amino acids would result. The initiation step is also of great importance in another respect, since 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 thus determines the rate at which the 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 initiate translation. This **initiator tRNA** always carries the amino acid methionine (in bacteria, a modified form of methionine—formylmethionine—is used) so that all newly made proteins have methionine as the first amino acid at their N-terminal end, the end of a protein that is synthesized first. This

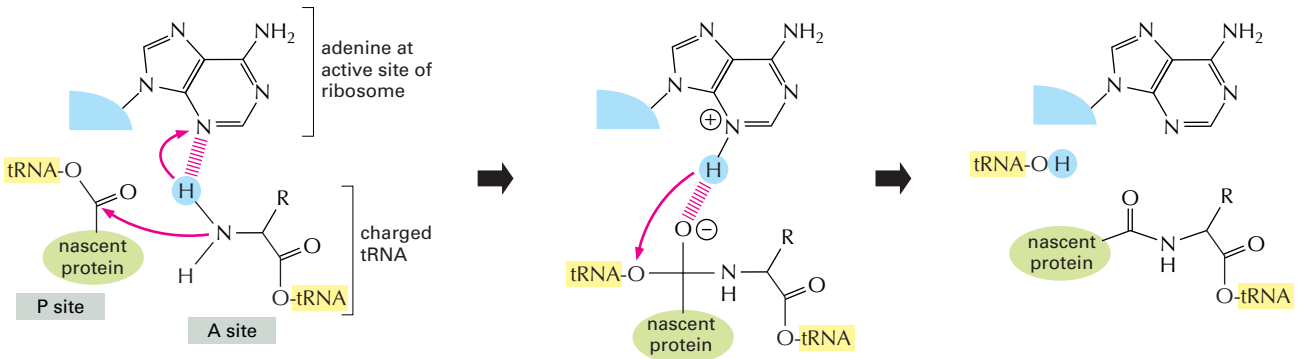
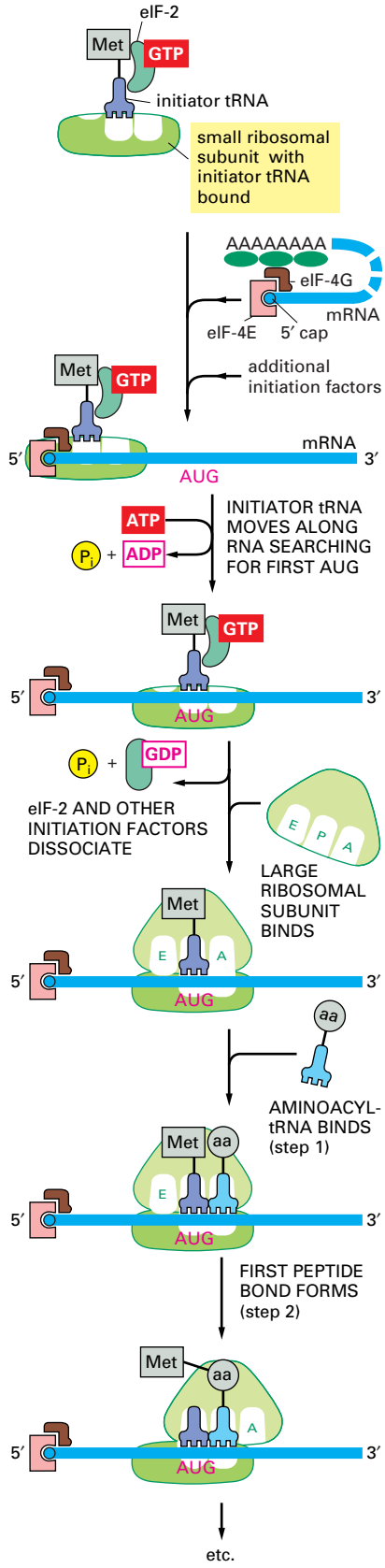


Figure 6–70 A possible reaction mechanism for the peptidyl transferase activity present in the large ribosomal subunit. The overall reaction is catalyzed by an active site in the 23S rRNA. In the first step of the proposed mechanism, the N3 of the active-site adenine abstracts a proton from the amino acid attached to the tRNA at the ribosome's A-site, allowing its amino nitrogen to attack the carboxyl group at the end of the growing peptide chain. In the next step this protonated adenine donates its hydrogen to the oxygen linked to the peptidyl-tRNA, causing this tRNA's release from the peptide chain. This leaves a polypeptide chain that is one amino acid longer than the starting reactants. The entire reaction cycle would then repeat with the next aminoacyl tRNA that enters the A-site. (Adapted from P. Nissen et al., *Science* 289:920–930, 2000.)

Figure 6–71 The initiation phase 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 (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.



methionine is usually removed later by a specific protease. The initiator tRNA has a nucleotide sequence distinct from that of the tRNA that normally carries methionine.

In eucaryotes, the initiator tRNA (which is coupled to methionine) is first loaded into the small ribosomal subunit along with additional proteins called **eucaryotic initiation factors**, or **eIFs** (Figure 6–71). 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 present. 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. This movement is facilitated by additional initiation factors that act as ATP-powered helicases, allowing the small subunit to scan 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 from the small ribosomal subunit to make way for the large ribosomal subunit to assemble with it and complete the ribosome. The initiator tRNA is now bound to the P-site, leaving the A-site vacant. Protein synthesis is therefore ready to begin with the addition of the next aminoacyl tRNA molecule (see Figure 6–71).

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 is quite different from the consensus recognition sequence, 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 tell 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 6–72). In contrast, a eucaryotic mRNA generally encodes only a single protein.

Stop Codons Mark the End of Translation

The end of the protein-coding message is signaled by the presence of one of three codons (UAA, UAG, or UGA) called *stop codons* (see Figure 6–50). These are not recognized by a tRNA and do not specify an amino acid, but instead signal

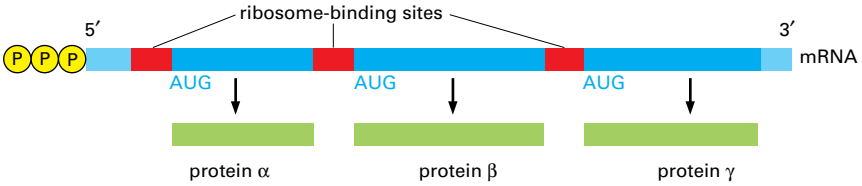


Figure 6–72 Structure of a typical bacterial mRNA molecule. Unlike eucaryotic ribosomes, which typically require a capped 5′ end, procaryotic ribosomes initiate transcription at ribosome-binding sites (Shine–Dalgarno sequences), which can be located anywhere along an mRNA molecule. This property of ribosomes permits bacteria to synthesize more than one type of protein from a single mRNA molecule.

to the ribosome to stop translation. Proteins known as *release factors* bind to any ribosome with a stop codon positioned in the A site, and this binding forces the peptidyl transferase in the ribosome to catalyze the addition of a water molecule instead of an amino acid to the peptidyl-tRNA (Figure 6–73). 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 another mRNA molecule to begin a new round of protein synthesis.

Release factors provide a dramatic 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) bears an uncanny resemblance to the shape and charge distribution of a tRNA molecule (Figure 6–74). This shape and charge mimicry allows the release factor to 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 6–68C). 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, because it is not complementary to any peptide structure, 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 structure to be useful to the cell, and later in this chapter we discuss how this folding occurs. First, however, we review 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. But even during this very short period, multiple initiations usually take place on each mRNA molecule being translated. As soon as the preceding ribosome has translated enough of the nucleotide sequence to move out of the way, the 5′ end of the mRNA is threaded into a new ribosome. The mRNA molecules being translated are therefore usually found in the form of *polyribosomes* (also known as *polysomes*), large cytoplasmic assemblies made up of several ribosomes spaced as close as 80 nucleotides apart along a single mRNA molecule (Figure 6–75). These multiple initiations mean that many more protein

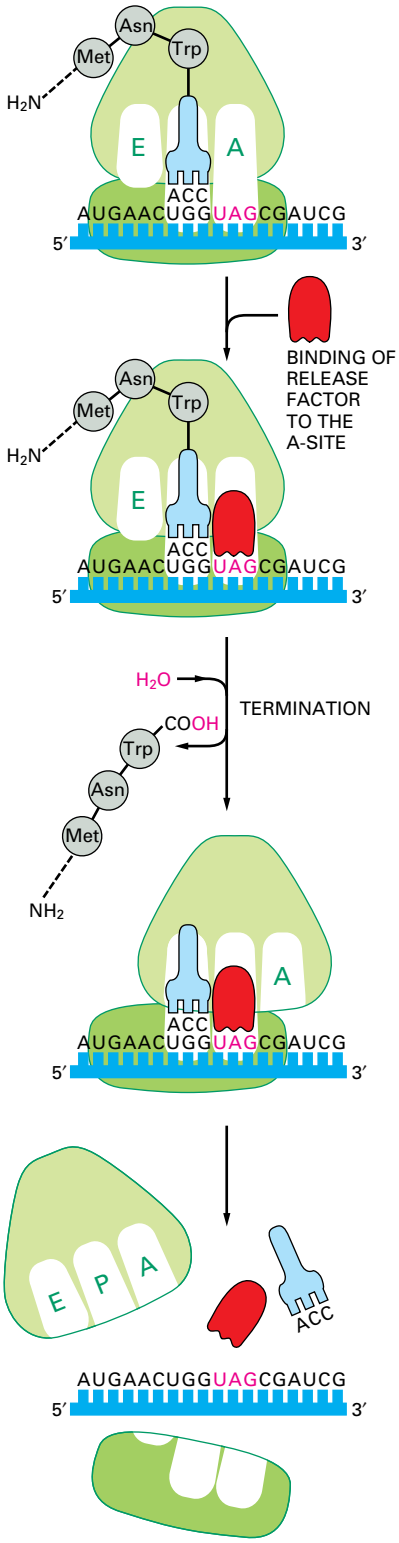


Figure 6–73 The final phase of protein synthesis. The binding of a release factor to an A-site bearing a stop codon terminates translation. The completed polypeptide is released and, after the action of a *ribosome recycling factor* (not shown), the ribosome dissociates into its two separate subunits.

Figure 6–74 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. © Elsevier.)

molecules can be made in a given time than would be possible if each had to be completed before the next could start.

Both bacteria and eucaryotes utilize polysomes, and both employ additional strategies to speed up the 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 6–75A); therefore, as soon as a ribosome dissociates, its two subunits are in an optimal position to reinitiate translation on the same mRNA molecule.

Quality-Control Mechanisms Operate at Many Stages of Translation

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 a time delay each time a new amino acid is added to a growing polypeptide chain, producing an overall

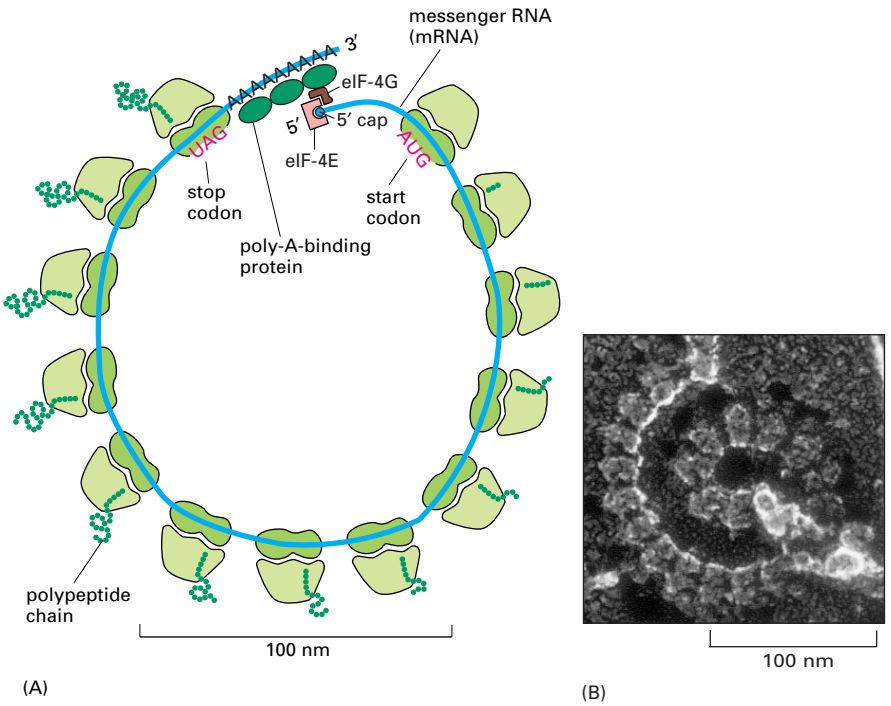
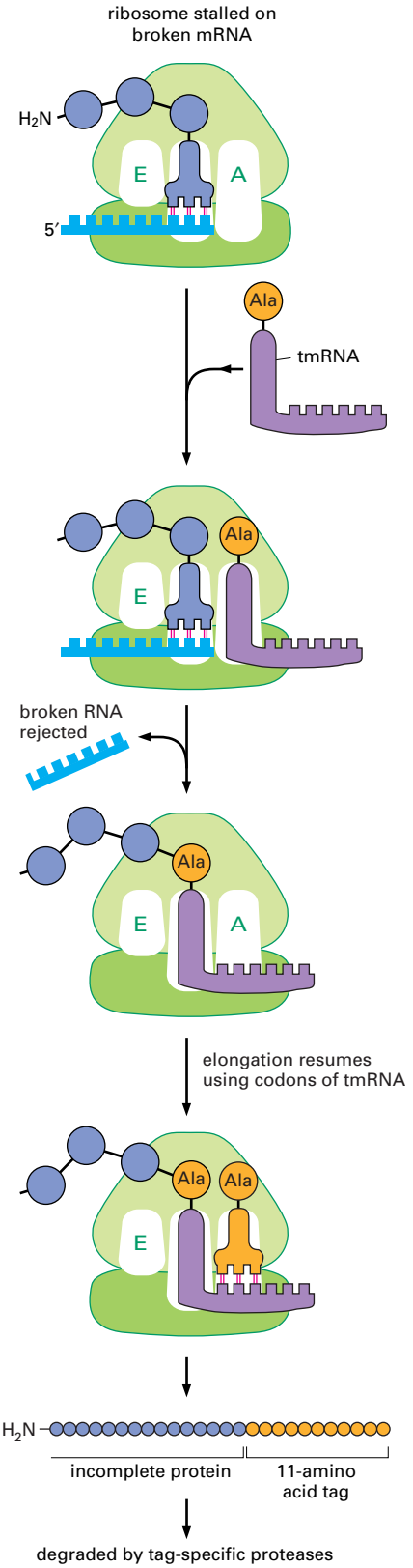


Figure 6–75 A polyribosome. (A) Schematic drawing showing how a series of ribosomes can simultaneously translate the same eucaryotic mRNA molecule. (B) Electron micrograph of a polyribosome from a eucaryotic cell. (B, courtesy of John Heuser.)

Figure 6–76 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 except that no codon is present to guide it. The ribosome then translates ten codons from the tmRNA, completing an 11-amino acid tag on the protein. This tag is recognized by proteases that then degrade the entire protein.



speed of translation of 20 amino acids incorporated per second in bacteria. Mutant bacteria with a specific alteration in their small ribosomal subunit 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, a price must 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 6–66). 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 (Figure 6–66). To be effective, these proofreading mechanisms must also remove an appreciable fraction of correct interactions; for this reason they are even more costly in energy than they might seem.

Other quality control mechanisms ensure that a eucaryotic mRNA molecule is complete before ribosomes even begin to translate it. Translating broken or partly processed mRNAs would be harmful to the cell, because truncated or otherwise aberrant proteins would be produced. In eucaryotes, we have seen that mRNA production involves not only transcription but also 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). An mRNA molecule that was intact when it left the nucleus can, however, become broken in the cytosol. To avoid translating such broken mRNA molecules, the 5′ cap and the poly-A tail are both recognized by the translation-initiation apparatus before translation begins (see Figures 6–71 and 6–75).

Bacteria solve the problem of incomplete mRNAs in an entirely different way. Not only are there no signals at the 3′ ends of bacterial mRNAs, but also, as we have seen, translation often begins before the synthesis of the transcript has been completed. When the bacterial ribosome translates to the end of an incomplete RNA, a special RNA (called *tmRNA*) enters the A-site of the ribosome and is itself translated; this adds a special 11 amino acid tag to the C terminus of the truncated protein that signals to proteases that the entire protein is to be degraded (Figure 6–76).

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 ancestry of all life on Earth. Although rare, there are exceptions to this code, and we discuss some of them in this section. 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) also show several deviations 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. 814).

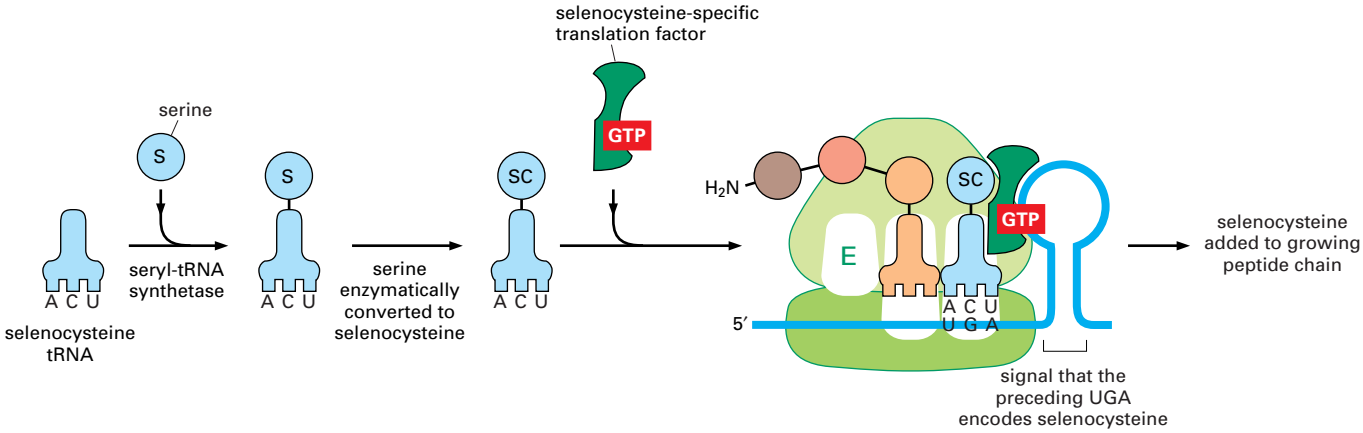


Figure 6–77 Incorporation of selenocysteine into a growing polypeptide chain. A specialized tRNA is charged with serine by the normal seryl-tRNA synthetase, and the serine is subsequently converted enzymatically to selenocysteine. A specific RNA structure in the mRNA (a stem and loop structure with a 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.

The type of deviation in the genetic code discussed above is "hardwired" into the organisms or the organelles in which it occurs. A different type of variation, sometimes called *translational 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 translational 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 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 is *translational frameshifting*. This type of recoding is commonly used by retroviruses, a large group of eucaryotic viruses, in which it allows more than one protein to be synthesized from a single mRNA. These viruses commonly 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). Such a virus needs many more copies of the Gag proteins than it does of the Pol proteins, and they achieve this quantitative adjustment by encoding the *pol* genes just after the *gag* genes but in a different reading frame. A stop codon at the end of the *gag* coding sequence can be bypassed on occasion by an intentional translational frameshift that occurs upstream of it. 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).

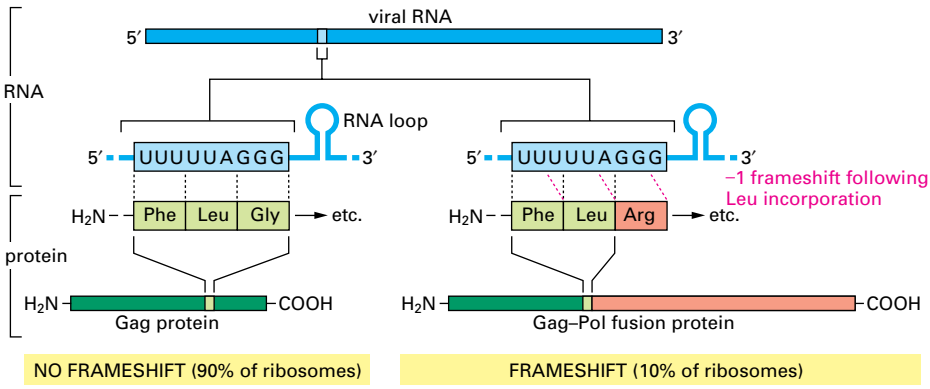


Figure 6–78 The translational frameshifting that produces the reverse transcriptase and integrase of a retrovirus. The viral reverse transcriptase and integrase are produced by proteolytic processing of a large protein (the Gag–Pol fusion protein) consisting of both the Gag and Pol amino acid sequences. The viral capsid proteins are produced by proteolytic processing of the more abundant Gag protein. Both the Gag and the Gag–Pol fusion proteins start identically, but the Gag protein terminates at an in-frame stop codon (not shown); the indicated frameshift bypasses this stop codon, allowing the synthesis of the longer Gag–Pol fusion protein. The frameshift occurs because 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 stem and loop structure that forms in the viral mRNA, as indicated in the figure. The sequence shown is from the human AIDS virus, HIV. (Adapted from T. Jacks et al., *Nature* 331:280–283, 1988.)

Many 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 act by inhibiting bacterial protein synthesis. Some of these drugs exploit the structural and functional differences between bacterial and eucaryotic ribosomes so as to interfere preferentially with the function of bacterial ribosomes. Thus some of these compounds can be taken in high doses without undue toxicity to humans. Because different antibiotics bind to different regions of bacterial ribosomes, they often inhibit different steps in the synthetic process. Some of the more common antibiotics of this kind are listed in Table 6–3 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 6–3 are useful for cell biological studies. Among the most commonly used drugs in such experimental studies 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.

Having described the translation process itself, we now discuss how its products—the proteins of the cell—fold into their correct three-dimensional conformations.

A Protein Begins to Fold While It Is 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

TABLE 6–3 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 initiation complex to chain-elongating ribosome and also causes miscoding
Chloramphenicol	blocks the peptidyl transferase reaction on ribosomes (step 2 in Figure 6–65)
Erythromycin	blocks the translocation reaction on ribosomes (step 3 in Figure 6–65)
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 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 6–65)
Anisomycin	blocks the peptidyl transferase reaction on ribosomes (step 2 in Figure 6–65)
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.

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 6–79).

The information needed for all of the protein maturation 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. 134).

Through many millions of years of evolutionary time, 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, as its polypeptide chain spins out of the ribosome starting from the N-terminal end. Experiments have demonstrated that once a protein domain in a multi-domain protein emerges from the ribosome, it forms a compact structure within a few seconds that contains most of the final secondary structure (α helices and β sheets) aligned in roughly the right way (Figure 6–80). For many protein domains, this unusually open and flexible structure, which is called a *molten globule*, is the starting point for a relatively slow process in which many side-chain adjustments occur that eventually form the correct tertiary structure. Nevertheless, because it takes several minutes to synthesize a protein of average size, a great deal of the folding process is complete by the time the ribosome releases the C-terminal end of a protein (Figure 6–81).

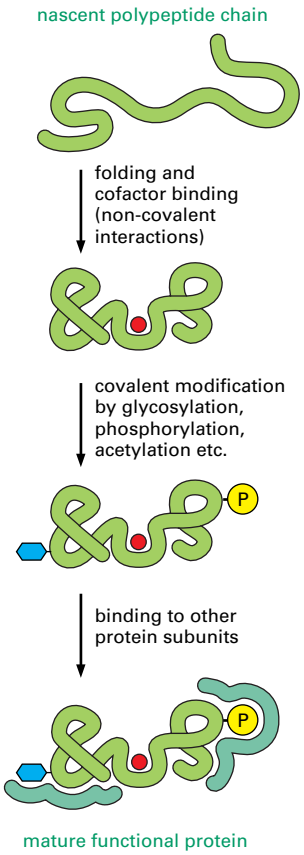


Figure 6–79 Steps in the creation of a functional protein. As indicated, translation of an mRNA sequence into an amino acid sequence on the ribosome is not the end of the process of forming a protein. To be useful to the cell, the completed polypeptide chain must fold correctly into its three-dimensional conformation, bind any cofactors required, and assemble with its partner protein chains (if any). These changes are driven by noncovalent bond formation. As indicated, many proteins also have covalent modifications made to selected amino acids. Although the most frequent of these are protein glycosylation and protein phosphorylation, more than 100 different types of covalent modifications are known (see, for example, Figure 4–35).

Molecular Chaperones Help Guide the Folding of Many Proteins

The folding of many proteins is made more efficient by a special class of proteins called **molecular chaperones**. The latter proteins are useful for cells because there are a variety of different paths that can be taken to convert the molten globule form of a protein to the protein's 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 that resets the folding process (Figure 6–82).

Molecular chaperones were first identified in bacteria when *E. coli* mutants that failed to allow bacteriophage lambda to replicate in them were studied.

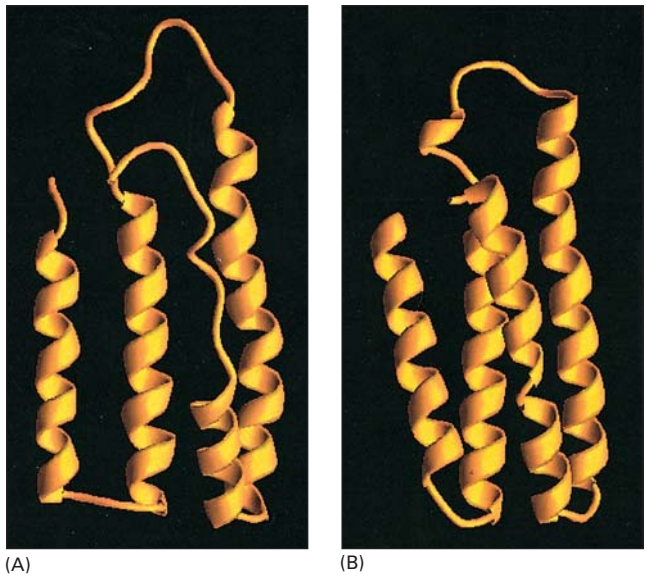


Figure 6–80 The structure of a molten globule. (A) A molten globule form of cytochrome b_{562} is more open and less highly ordered than the final folded form of the protein, shown in (B). Note that the molten globule contains most of the secondary structure of the final form, although the ends of the α helices are frayed 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.)

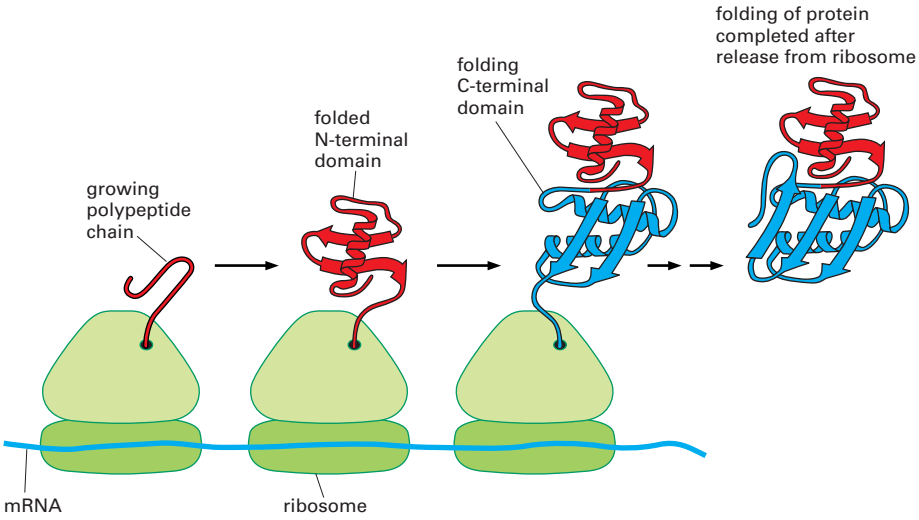


Figure 6–81 The co-translational folding of a protein. A growing polypeptide chain is shown acquiring its secondary and tertiary structure as it emerges from a ribosome. The N-terminal domain folds first, while the C-terminal domain is still being synthesized. In this case, the protein has not yet achieved its final conformation by the time it is released from the ribosome. (Modified from A.N. Federov and T.O. Baldwin, *J. Biol. Chem*. 272:32715–32718, 1997.)

These mutant cells produce slightly altered versions of the chaperone machinery, and as a result they are defective in specific steps in the assembly of the viral proteins. The molecular chaperones are included among the *heat-shock proteins* (hence their designation as *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 any increase in misfolded proteins (such as those produced by elevated temperatures) by boosting the synthesis of the chaperones that help these proteins refold.

Eucaryotic cells have at least two major families of molecular chaperones—known as 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.

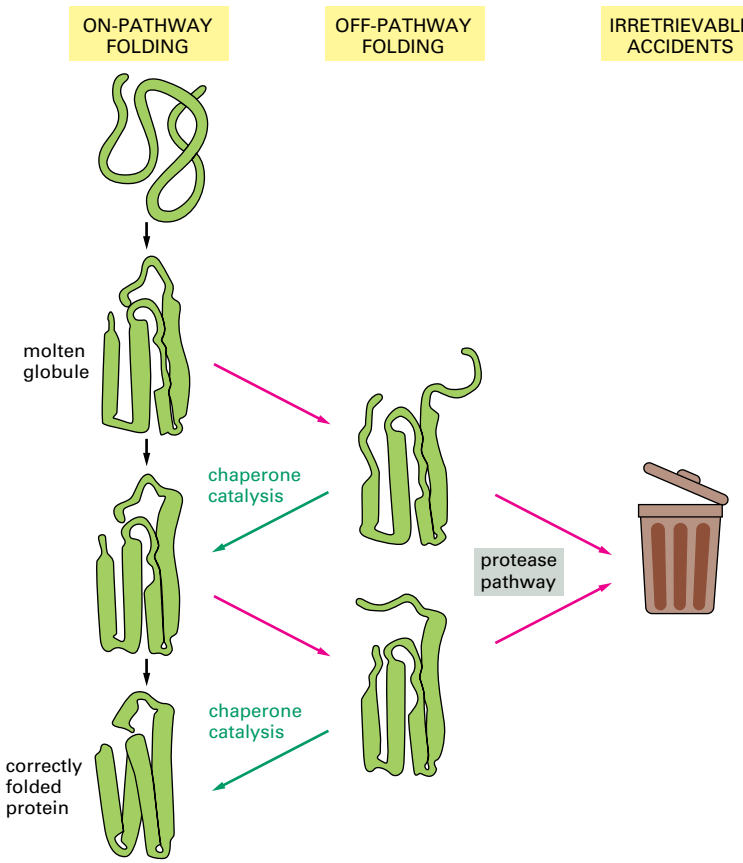


Figure 6–82 A current view of protein folding. Each domain of a newly synthesized protein rapidly attains a "molten globule" state. Subsequent folding occurs more slowly and by multiple pathways, often involving the help of a molecular chaperone. Some molecules may still fail to fold correctly; as explained shortly, these are recognized and degraded by specific proteases.

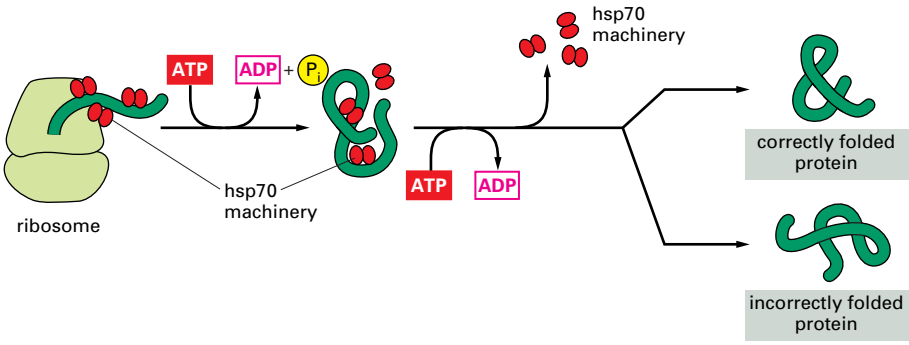


Figure 6–83 The hsp70 family of molecular chaperones. These proteins act early, recognizing a small stretch of hydrophobic amino acids on a protein's surface. Aided by a set of smaller hsp40 proteins, an hsp70 monomer binds to its target protein and then hydrolyzes a molecule of ATP to ADP, undergoing a conformational change that causes the hsp70 to clamp down very tightly on the target. After the hsp40 dissociates, the dissociation of the hsp70 protein is induced by the rapid re-binding of ATP after ADP release. Repeated cycles of hsp protein binding and release help the target protein to refold, as schematically illustrated in Figure 6–82.

The hsp60-like and hsp70 proteins each work with their own small set of associated proteins when they help other proteins to fold. They share an affinity for the exposed hydrophobic patches on incompletely folded proteins, and they hydrolyze ATP, often binding and releasing their protein 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 6–83). In contrast, hsp60-like proteins form a large barrel-shaped structure that acts later in a protein's life, after it has been fully synthesized. This type of chaperone 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 6–84).

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. It is this type of experiment that shows that the hsp70 proteins act first, beginning when a protein is still being synthesized on a ribosome, and that the

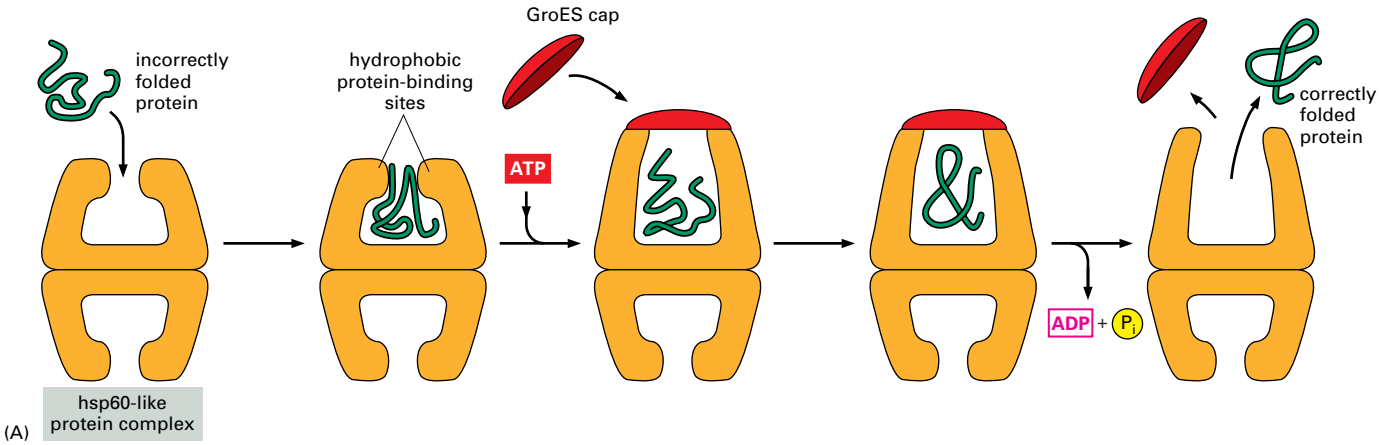


Figure 6–84 The structure and function of the hsp60 family of molecular chaperones. (A) The catalysis of protein refolding. As indicated, a misfolded protein is initially captured by hydrophobic interactions along one rim of the barrel. The subsequent binding of ATP plus a protein cap increases the diameter 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 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, TCP-1 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. Bukace and A.L. Horwich, *Cell* 92:351–366, 1998.)

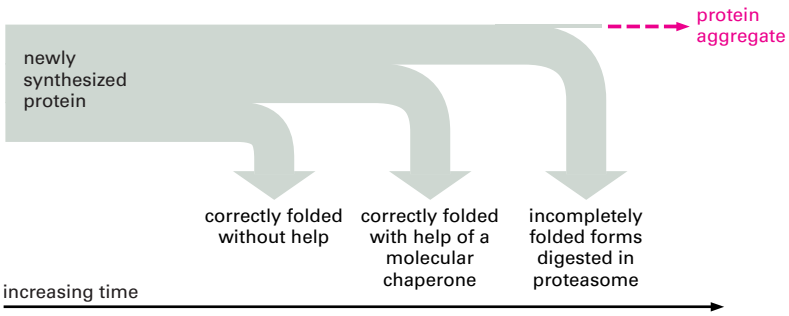


Figure 6–85 The cellular mechanisms that monitor protein quality after protein synthesis. As indicated, a newly synthesized protein sometimes folds correctly and assembles with its partners on its own, in which case it is left alone. Incompletely folded proteins are helped to refold by molecular chaparones: first by a family of hsp70 proteins, and if this fails, then by hsp60-like proteins. In both cases the client proteins are recognized by an abnormally exposed patch of hydrophobic amino acids on their surface. These processes compete with a different system that recognizes an abnormally exposed patch and transfers the protein that contains it to a proteasome for complete destruction. The combination 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 and precipitate the entire mass out of solution.

hsp60-like proteins are called into play only later to help in folding completed proteins. However, the same experiments reveal that only a subset of the newly synthesized proteins becomes involved: perhaps 20% of all proteins with the hsp70 and 10% with the hsp60-like molecular chaperones. How are these proteins selected for this ATP-catalyzed refolding?

Before answering, we need to pause to consider the post-translational fate of proteins more broadly. A protein that has a sizable exposed patch of hydrophobic amino acids on its surface is usually 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, precipitating out of solution. We shall see that, in rare cases, such aggregates do form and cause severe human diseases. But in the vast majority of cells, powerful protein quality control mechanisms prevent such disasters.

Given this background, it is not surprising that cells have evolved elaborate mechanisms that recognize and remove the hydrophobic patches on proteins. 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 are therefore bypassed by chaperones.

Figure 6–85 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 Degrades a Substantial Fraction of the Newly Synthesized Proteins in Cells

Cells quickly remove the failures of their translation processes. Recent experiments suggest that as many as one-third of the newly made polypeptide chains are selected for rapid degradation as a result of the protein quality control mechanisms just described. The final disposal apparatus in eucaryotes is the **proteasome**, an abundant ATP-dependent protease that constitutes nearly 1% of cellular protein. Present in many copies dispersed throughout the cytosol and the nucleus, the proteasome also targets proteins of the endoplasmic reticulum (ER): those proteins that fail either to fold or to be assembled properly after they enter the ER are detected by an ER-based surveillance system that *retrotranslocate* 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 cylindrical

stack of four heptameric rings. Some of these subunits are distinct proteases whose active sites face the cylinder's inner chamber (Figure 6–86A). Each end of the cylinder is normally associated with a large protein complex (the 19S cap) containing approximately 20 distinct polypeptides (Figure 6–86B). The cap subunits include at least six proteins that hydrolyze ATP; located near the edge of the cylinder, these ATPases are thought to unfold the proteins to be digested and move them into the interior chamber for proteolysis. 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 act as regulated "gates" at the entrances to the inner proteolytic chamber, being also 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 multiple copies of a small protein called *ubiquitin* (Figure 6–87A). Ubiquitin exists in cells either free or covalently linked to a huge variety of intracellular proteins. For most of these proteins, this tagging by ubiquitin results in their destruction by the proteasome.

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 ubiquitin that is transferred to one of a set of ubiquitin-conjugating (E2) enzymes. 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 in protein substrates, helping E2 to form a *multiubiquitin* 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, producing a linear series of ubiquitin–ubiquitin conjugates (Figure 6–87B). It is this multiubiquitin 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 for degradation distinct subsets of intracellular proteins that bear these signals.

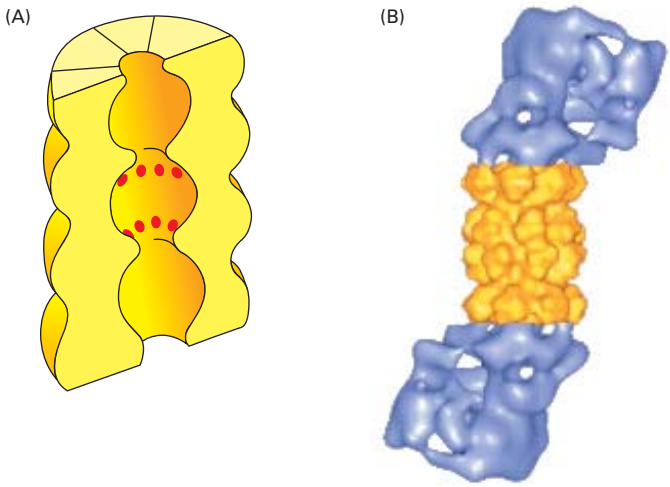


Figure 6–86 The proteasome. (A) A cut-away view of the structure of the central 20S cylinder, as determined by x-ray crystallography, with the active sites of the proteases indicated by *red dots*. (B) The structure of the entire proteasome, in which the central cylinder *(yellow)* is supplemented by a 19S cap *(blue)* at each end, whose structure has been determined by computer processing of electron microscope images. The complex cap structure selectively binds those proteins that have been marked for destruction; it then uses ATP hydrolysis to unfold their polypeptide chains and feed them into the inner chamber of the 20S cylinder for digestion to short peptides. (B, from W. Baumeister et al., *Cell* 92:367–380, 1998. © Elsevier.)

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 some 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.

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 half-lives on specific normal proteins whose concentrations must change promptly with alterations

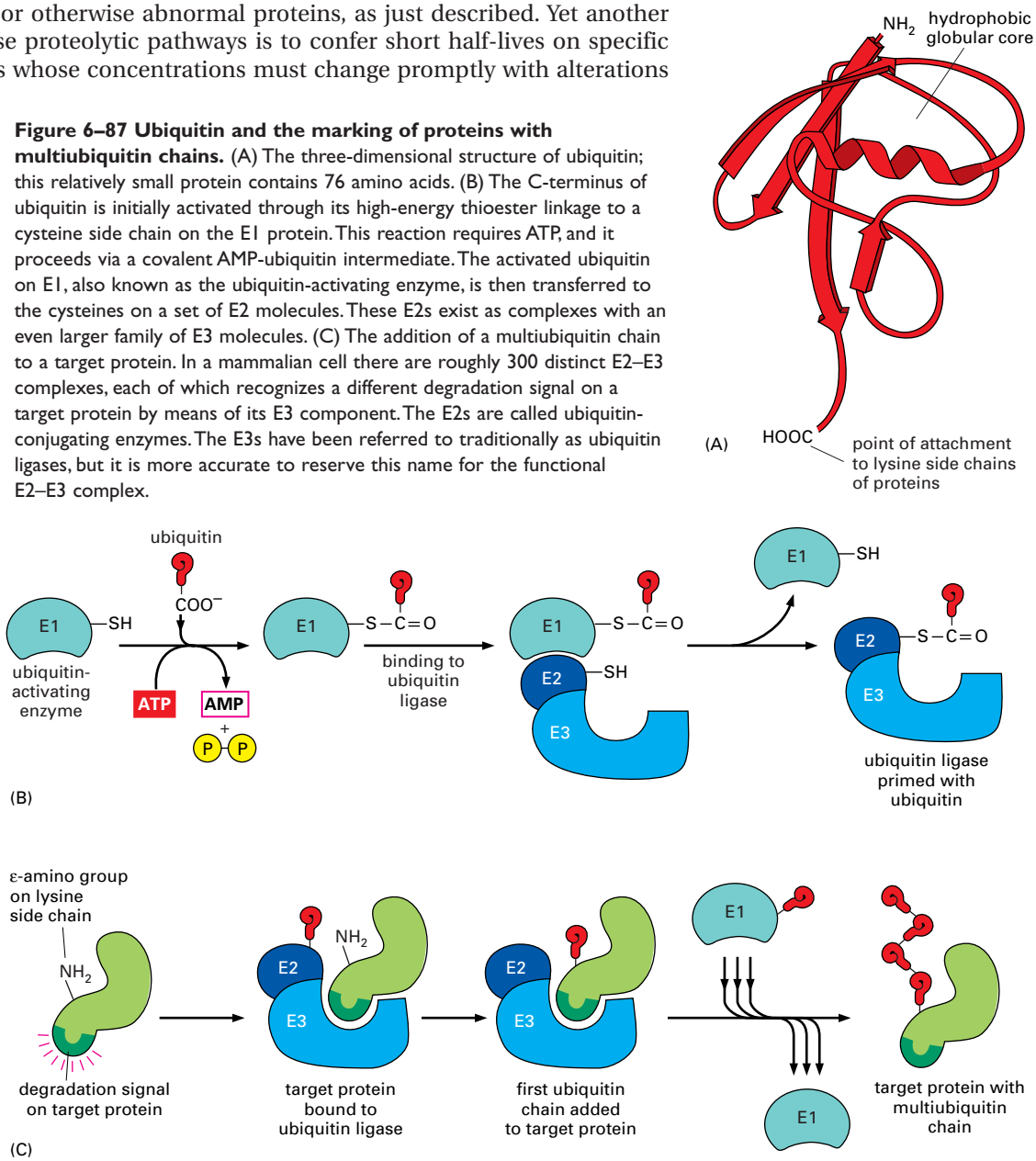


Figure 6–87 Ubiquitin and the marking of proteins with multiubiquitin chains. (A) The three-dimensional structure of ubiquitin; this relatively small protein contains 76 amino acids. (B) The C-terminus of ubiquitin is initially activated through its high-energy thioester linkage to a cysteine side chain 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 multiubiquitin chain to a target protein. In a mammalian cell there are roughly 300 distinct E2–E3 complexes, each of which recognizes a different degradation signal on a target protein by means of its 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.

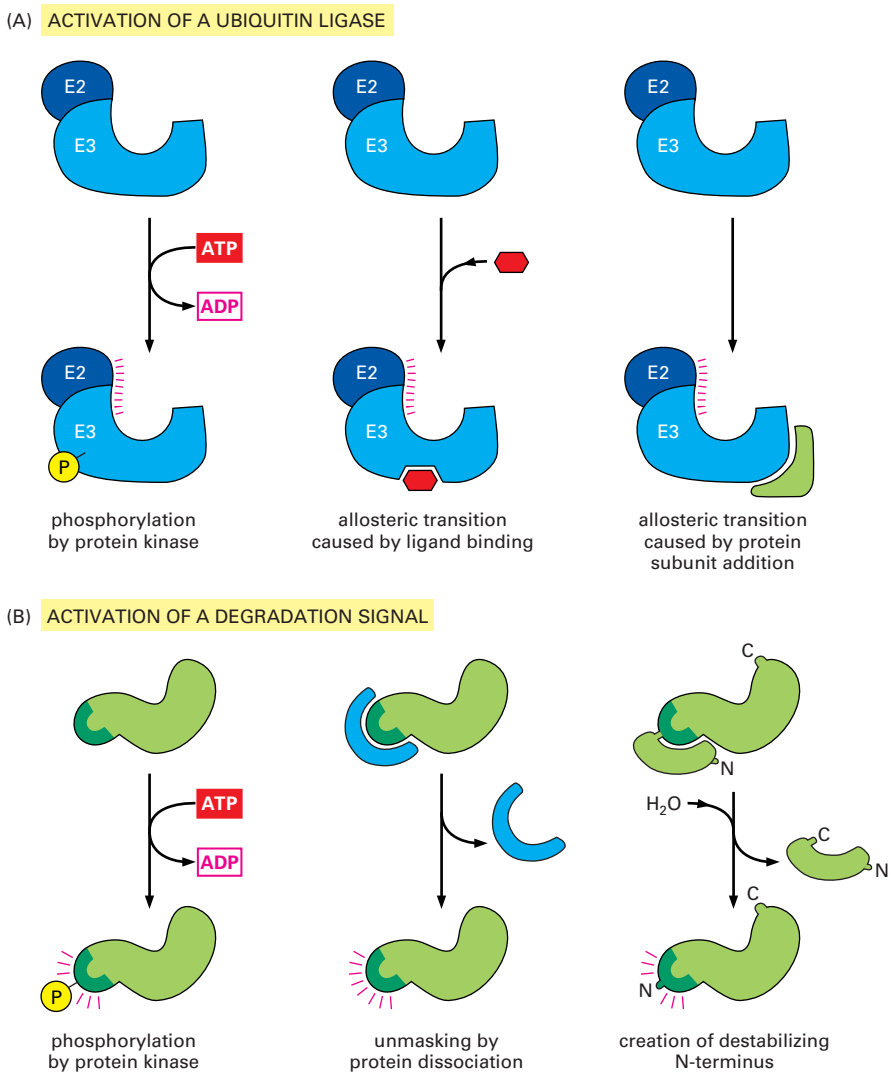


Figure 6–88 Two general ways of inducing the degradation of a specific protein. (A) Activation of a specific E3 molecule creates a new ubiquitin ligase. (B) Creation of an exposed degradation signal in the protein to be degraded. This signal binds a ubiquitin ligase, causing the addition of a multiubiquitin 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.

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? A variety of mechanisms are known, as illustrated through specific examples later in this book. In one general class of mechanism (Figure 6–88A), 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 17–20).

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 a single cleavage of a 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 6–88B).

The N-terminal type of degradation signal arises because of the "N-end rule," which relates the half-life 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 in the yeast-type N-end rule. Therefore, it was initially unclear how N-end rule substrates form *in vivo*. However, it has recently been discovered that a subunit of cohesin, a protein complex that holds sister chromatids together, is cleaved by a site-specific protease at the metaphase–anaphase transition. This cell-cycle-regulated cleavage allows separation of the sister chromatids and leads to the completion of mitosis (see Figure 17–26). 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

When all of a cell's protein quality controls fail, large protein aggregates tend to accumulate in the affected cell (see Figure 6–85). Some of these aggregates, by adsorbing critical macromolecules to them, can severely damage cells and even cause cell death. The protein aggregates released from dead cells tend to 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. Not surprisingly, therefore, protein aggregates primarily cause diseases of neurodegeneration. 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 6–89C) tends to be highly resistant to proteolysis. This resistance presumably explains why this structure is observed in so many of the neurological disorders caused by protein aggregates, where it produces abnormally staining deposits known as *amyloid*.

One particular variety of these diseases has attained special notoriety. These are the **prion diseases**. Unlike Huntington's or Alzheimer's disease, a prion disease 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 function is not known. However, PrP has the unfortunate property of being convertible to a very special abnormal conformation (Figure 6–89A). This conformation not only forms protease-resistant, cross-beta filaments; it also is "infectious" because it converts normally folded molecules of PrP to the same form. This property creates a positive feedback loop that propagates the abnormal form of PrP, called PrP* (Figure 6–89B) and thereby allows PrP to spread rapidly from cell to cell in the brain, causing the death of both

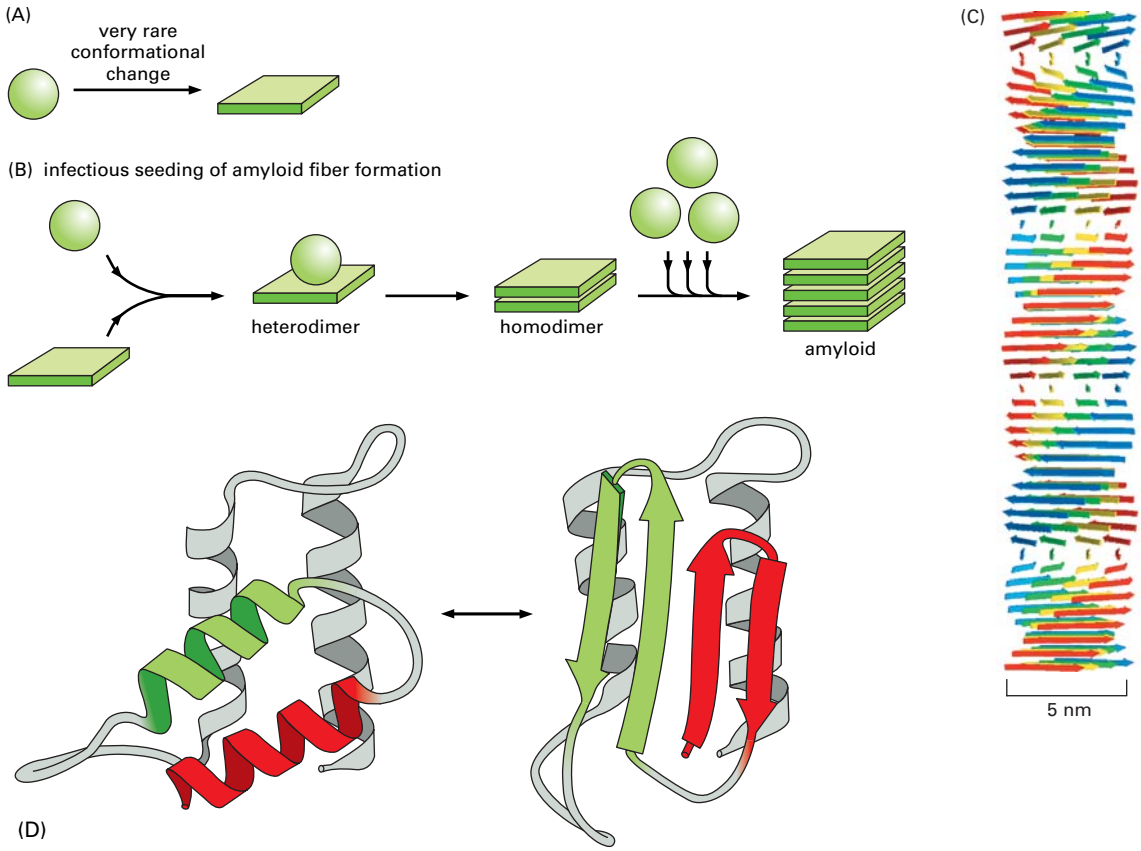


Figure 6–89 Protein aggregates that cause human disease. (A) Schematic illustration of the type of conformational change in a protein that produces material for a cross-beta filament. (B) Diagram illustrating the self-infectious nature of the protein aggregation that is central to prion diseases. PrP 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 because this structure is not infectious in this way, it cannot spread from one animal to another. (C) Drawing of a cross-beta filament, a common type of protease-resistant protein aggregate found in a variety of 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; D, adapted from S.B. Prusiner, *Trends Biochem. Sci.* 21:482–487, 1996.)

animals and humans. It can be dangerous to eat the tissues of animals that contain PrP*, as witnessed most recently by the spread of BSE (commonly referred to as the "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, a similar transformation has been discovered to be the cause of an otherwise mysterious "protein-only inheritance" observed in yeast cells.

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 6–90). The final level of a properly folded protein in a cell therefore depends upon the efficiency with which each of the many steps is performed.

We discuss in Chapter 7 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 Fig-

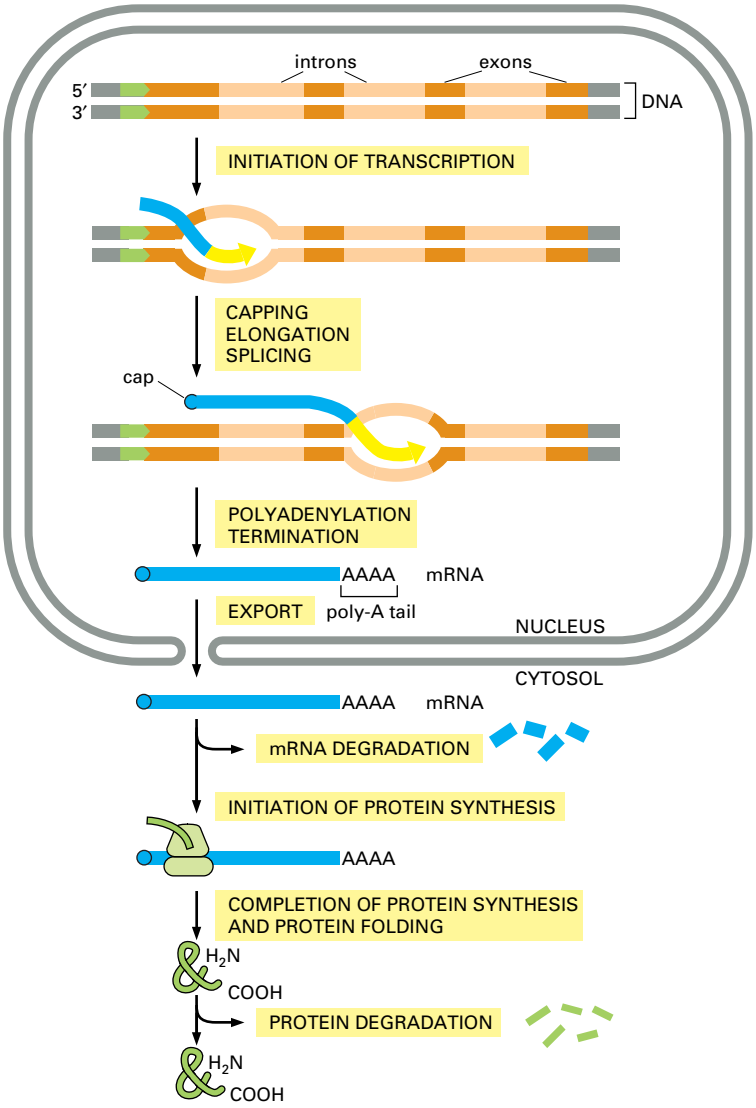


Figure 6–90 The production of a protein by a eucaryotic cell. The final level of each protein in a eucaryotic cell depends upon the efficiency of each step depicted.

ure 6–90) could be regulated by the cell for each individual protein. However, as we shall see in Chapter 7, 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 the elongation phase of protein synthesis. During this phase, aminoacyl tRNAs—each bearing a specific amino acid bind sequentially to the appropriate codon in mRNA by forming complementary base pairs with the tRNA anticodon. Each amino acid is added to the C-terminal end of the growing polypeptide by means of a cycle of three sequential

steps: aminoacyl-tRNA binding, followed by peptide bond formation, followed by ribosome translocation. The mRNA molecule progresses codon by codon through the ribosome in the 5′-to-3′ direction until one of three stop codons is reached. 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 providing the peptidyl transferase enzyme activity 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 a highly 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 multiubiquitin chain is recognized by the cap on a proteasome to move 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 lifetime 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

To fully understand the processes occurring in present-day living cells, we need to consider how they arose in evolution. The most fundamental of all such problems is the expression of hereditary information, which today requires extraordinarily complex machinery and proceeds from DNA to protein through an RNA intermediate. How did this machinery arise? One view is that an *RNA world* existed on Earth before modern cells arose (Figure 6–91). 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.

In this section we outline 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 Autocatalysis

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. But life requires molecules that possess a crucial property: the ability to catalyze reactions that lead, directly or indirectly, to the production of more molecules like themselves. Catalysts with this special self-promoting property can use raw materials to

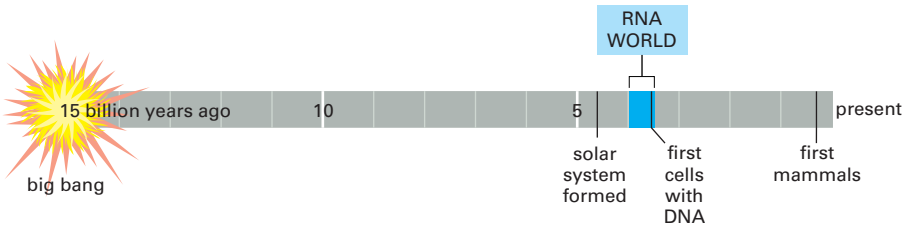


Figure 6–91 Time line for the universe, suggesting the early existence of an RNA world of living systems.

reproduce themselves and thereby divert these same materials from the production of other substances. But what molecules could have had such autocatalytic properties in early cells? 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. But, although polypeptides are versatile as catalysts, there is no known way in which one such molecule 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 exact 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 it be catalyzed before proteins with the appropriate enzymatic specificity existed? The beginnings of an answer to this question were obtained in 1982, when it was discovered that RNA molecules themselves can act as catalysts. We have seen in this chapter, for example, that a molecule of RNA is the catalyst for the peptidyl transferase reaction that takes place on the ribosome. The unique potential of RNA molecules to act both as information carrier and as catalyst forms the basis of the RNA world hypothesis.

RNA therefore has all the properties required of a molecule that could catalyze its own synthesis (Figure 6–92). Although self-replicating systems of RNA molecules have not been found in nature, scientists are hopeful 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 suggest that such a scenario is possible.

A Pre-RNA World Probably Predates the RNA World

Although RNA seems well suited to form the basis for a self-replicating set of biochemical catalysts, it is unlikely that RNA was the first kind of molecule to do so.

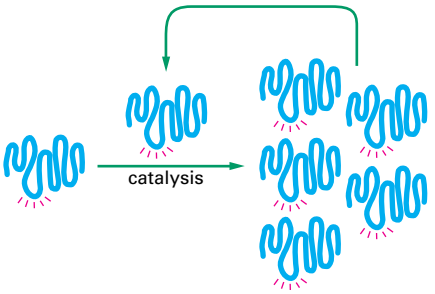


Figure 6–92 An RNA molecule that can catalyze its own synthesis. This hypothetical process would require catalysis of the production of both a second RNA strand of complementary nucleotide sequence and the use of this second RNA molecule as a template to form many molecules of RNA with the original sequence. The *red* rays represent the active site of this hypothetical RNA enzyme.

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 form in the face of a set of competing reactions, including hydrolysis, 2′ to 5′ linkages, 5′ to 5′ linkages, and so on. 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 6–93). 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" make them better candidates than RNA itself for the first biopolymers on Earth that had both information storage capacity and catalytic activity.

The transition between the pre-RNA world and the RNA world would have occurred through the synthesis of RNA using one of these simpler compounds as both template and catalyst. The plausibility of this scheme is supported by laboratory experiments showing that one of these simpler forms (PNA—see Figure 6–93) can act as a template for the synthesis of complementary RNA molecules, because the overall geometry of the bases is similar in the two molecules. Presumably, pre-RNA polymers also catalyzed the formation of ribonucleotide precursors from simpler molecules. Once the first RNA molecules had been produced, they could have diversified gradually to take over the functions originally carried out by the pre-RNA polymers, leading eventually to the postulated RNA world.

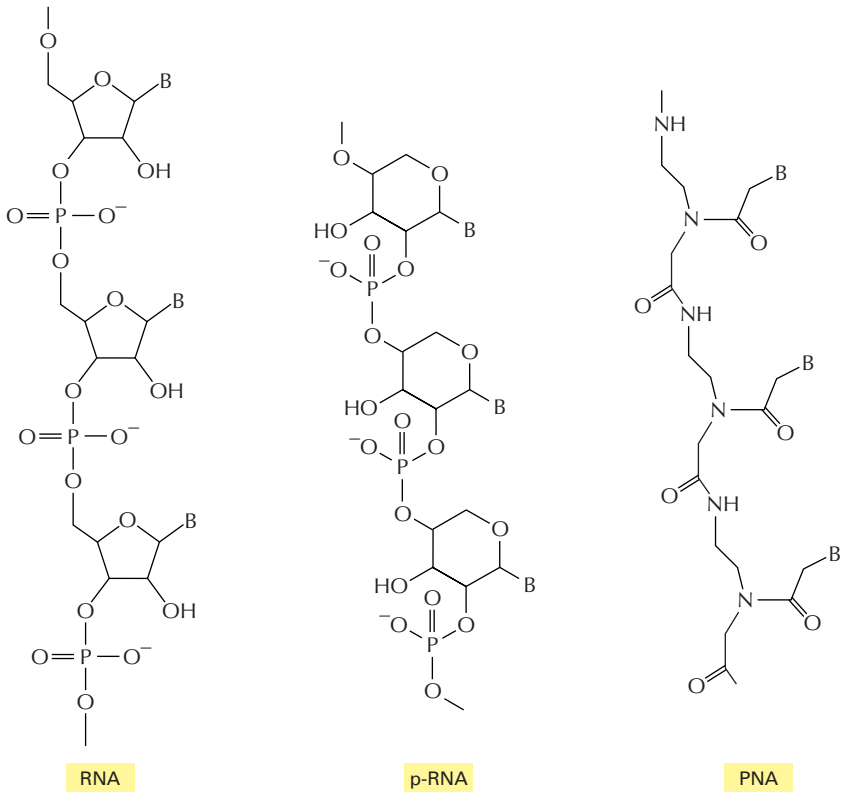


Figure 6–93 Structures of RNA and two related information-carrying polymers. In each case, B indicates the positions of purine and pyrimidine bases. The polymer p-RNA (pyranosyl-RNA) is RNA in which the furanose (five-membered ring) form of ribose has been replaced by the pyranose (six-membered ring) form. In PNA (peptide nucleic acid), the ribose phosphate backbone of RNA has been replaced by the peptide backbone found in proteins. Like RNA, both p-RNA and PNA can form double helices through complementary base-pairing, and each could therefore in principle serve as a template for its own synthesis (see Figure 6–92).

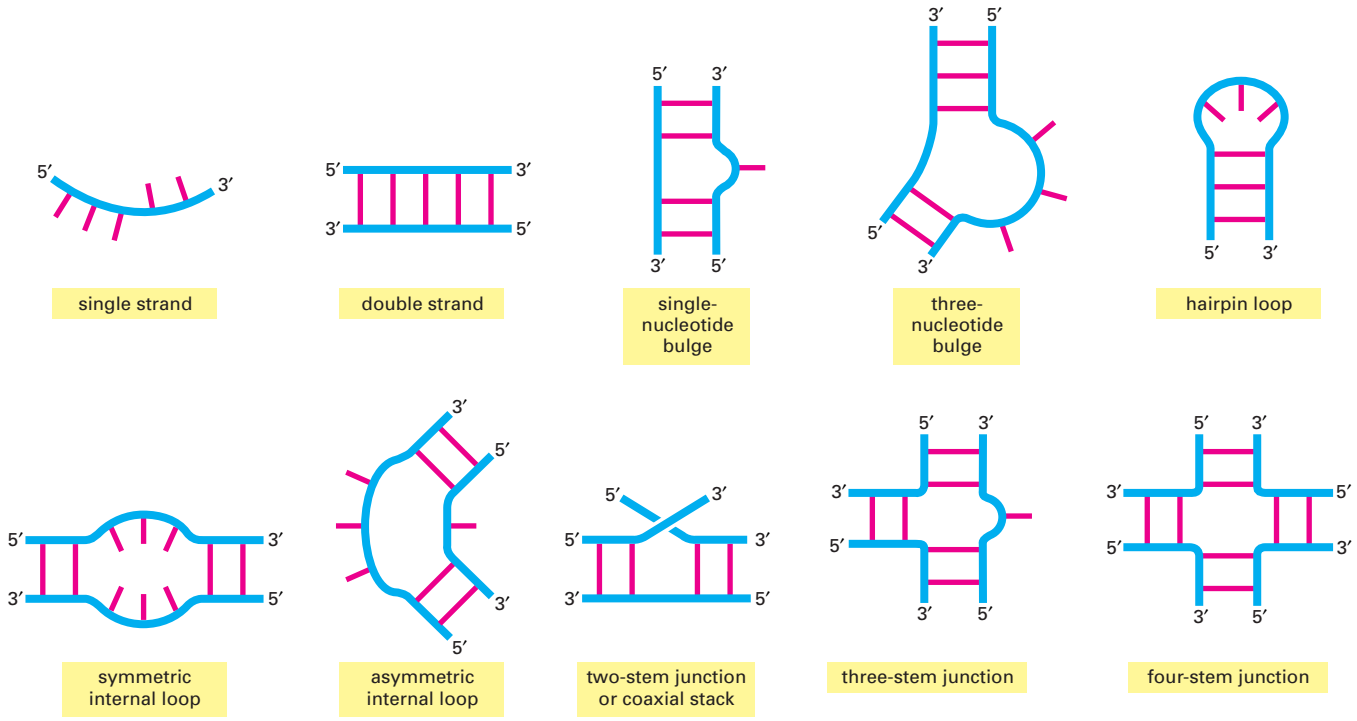


Figure 6–94 Common elements of RNA secondary structure. Conventional, complementary base-pairing interactions are indicated by *red* "rungs" in double-helical portions of the RNA.

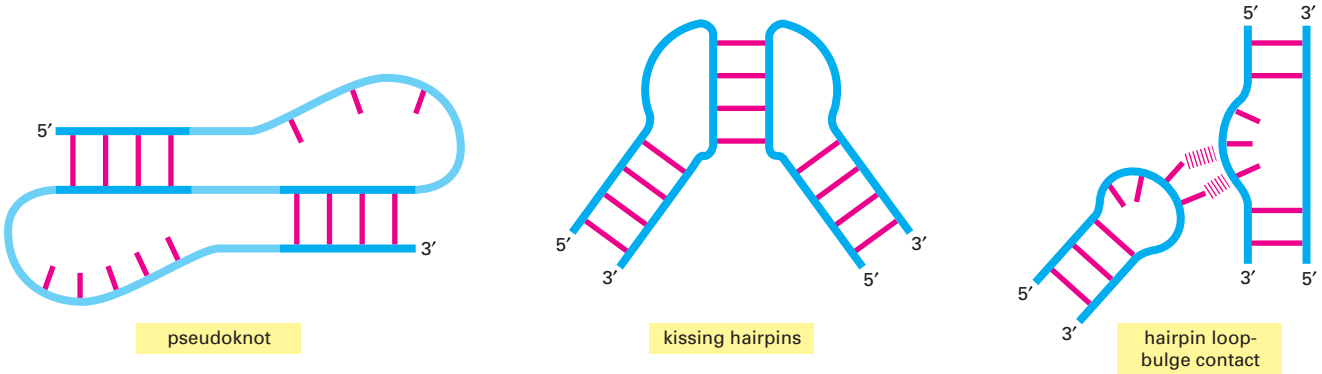
Single-stranded RNA Molecules Can Fold into Highly Elaborate Structures

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 6–67). Comparisons of many RNA structures have revealed conserved motifs, short structural elements that are used over and over again as parts of larger structures. Some of these RNA secondary structural motifs are illustrated in Figure 6–94. In addition, a few common examples of more complex and often longer-range interactions, known as RNA tertiary interactions, are shown in Figure 6–95.

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 6–96). 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 can be accounted for solely by the limited chemical groups of the polynucleotide chain.

Relatively few catalytic RNAs 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 6–97). Experiments of this type

Figure 6–95 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.



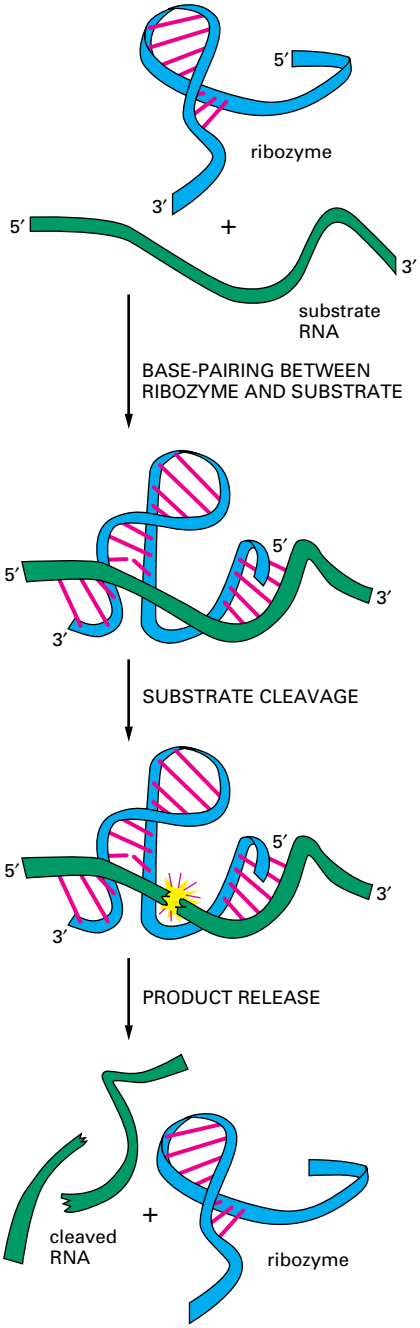


Figure 6–96 *(above)* A ribozyme. This simple RNA molecule catalyzes the cleavage of a second RNA at a specific site. This ribozyme is found embedded in larger RNA genomes—called viroids—which infect plants. The cleavage, which occurs in nature at a distant location on the same RNA molecule that contains the ribozyme, is a step in the replication of the viroid genome. Although not shown in the figure, the reaction requires a molecule of Mg positioned at the active site. (Adapted from T.R. Cech and O.C. Uhlenbeck, *Nature* 372:39–40, 1994.)

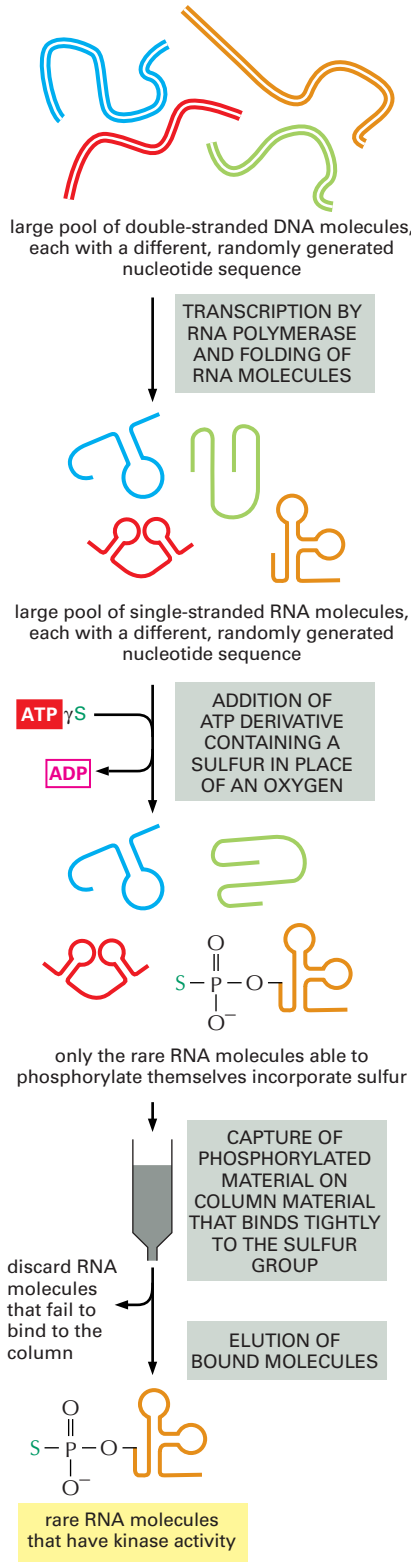


Figure 6–97 *(left)* *In vitro* selection of a synthetic ribozyme. Beginning with a large pool of nucleic acid molecules synthesized in the laboratory, those rare RNA molecules that possess a specified catalytic activity can be isolated and studied. Although a specific example (that of an autophosphorylating ribozyme) is shown, variations of this procedure have been used to generate many of the ribozymes listed in Table 6–4. During the autophosphorylation step, the RNA molecules are 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 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.)

TABLE 6–4 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; 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
Amide bond cleavage	*in vitro* selected RNA
Glycosidic bond formation	*in vitro* selected RNA
Porphyrin metalation	*in vitro* selected RNA

have created RNAs that can catalyze a wide variety of biochemical reactions (Table 6–4), and suggest that the main difference between protein enzymes and ribozymes lies in their maximum reaction speed, rather than in the diversity of the reactions that they can catalyze.

Like proteins, RNAs can undergo allosteric conformational changes, either in response to small molecules or to other RNAs. One artificially created ribozyme can exist in two entirely different conformations, each with a different catalytic activity (Figure 6–98). Moreover, the structure and function of the rRNAs in the ribosome alone have made it clear that RNA is an enormously versatile molecule. It is therefore easy to imagine that an RNA world could reach a 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.

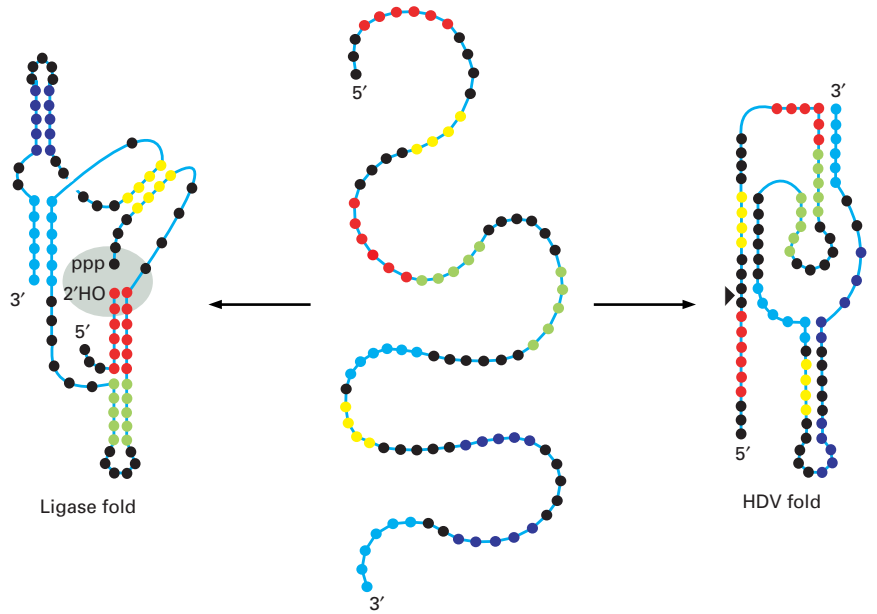


Figure 6–98 An RNA molecule that folds into two different ribozymes. This 88-nucleotide RNA, created in the laboratory, can fold into a ribozyme that carries out a self-ligation reaction *(left)* or a self-cleavage reaction *(right)*. The ligation reaction forms a 2′,5′ phosphodiester linkage with the 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.)

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. (This ribozyme activity has been directly demonstrated *in vitro*, albeit in a rudimentary form that can only synthesize moderate lengths of RNA.) 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 6–99). 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, may evolve into a cooperative system that replicates with unusually great efficiency.

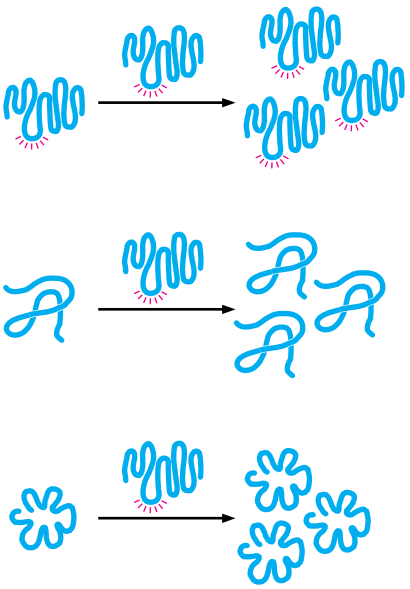


Figure 6–99 A family of mutually supportive RNA molecules, one catalyzing the reproduction of the others.

One of the crucial events leading to the formation of effective self-replicating systems must have been the development of individual compartments. For example, a set of mutually beneficial RNAs (such as those of Figure 6–99) could replicate themselves only if all the RNAs were to remain in the neighborhood of the RNA that is specialized for templated polymerization. Moreover, if these RNAs were free to diffuse among a large population of other RNA molecules, they could be co-opted by other replicating systems, which would then compete with the original RNA system for raw materials. Selection of a set of RNA molecules according to the quality of the self-replicating systems they generated could not occur efficiently until some form of compartment evolved to contain them and thereby make them available only to the RNA that had generated 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 *amphipathic*, 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. Amphipathic molecules of appropriate shape spontaneously aggregate to form *bilayers*, creating small closed vesicles whose aqueous contents are isolated from the external medium (Figure 6–100). 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 amphipathic molecules—mainly phospholipids—in this configuration; we discuss these molecules in detail in Chapter 10.

Presumably, the first membrane-bounded cells were formed by the spontaneous assembly of a set of amphipathic molecules, enclosing a self-replicating mixture of RNA (or pre-RNA) and other molecules. It is not clear at what point in the evolution of biological catalysts this first occurred. In any case, once RNA molecules were sealed within a closed membrane, they could begin to evolve in earnest as carriers of genetic instructions: they 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.

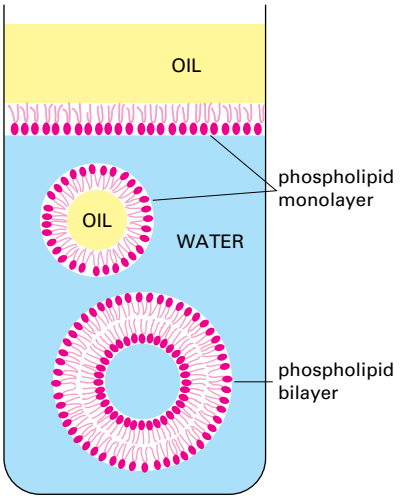


Figure 6–100 Formation of membrane by phospholipids. Because these molecules have hydrophilic heads and lipophilic tails, they align themselves at an oil/water interface with their heads in the water and their tails in the oil. In the water they associate to form closed bilayer vesicles in which the lipophilic tails are in contact with one another and the hydrophilic heads are exposed to the water.

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. Although we can only speculate on the origins of protein synthesis and the genetic code, several experimental approaches have provided possible scenarios.

In vitro RNA selection experiments of the type summarized previously in Figure 6–97 have produced RNA molecules that can bind tightly to amino acids. The nucleotide sequences of these RNAs often contain a disproportionately high frequency of codons for the amino acid that is recognized. For example, RNA molecules that bind selectively to arginine have a preponderance of Arg codons and those that bind tyrosine have a preponderance of Tyr codons. This correlation is not perfect for all the amino acids, and its interpretation is controversial, but it raises the possibility that a limited genetic code could have arisen from the direct association of amino acids with specific sequences of RNA, with RNAs serving as a crude template to direct the non-random polymerization of a few different amino acids. In the RNA world described previously, any RNA that helped guide the synthesis of a useful polypeptide would have a great advantage in the evolutionary struggle for survival.

In present-day cells, tRNA adaptors are used to match amino acids to codons, and proteins catalyze tRNA aminoacylation. However, ribozymes created in the laboratory can perform specific tRNA aminoacylation reactions, so it is plausible that tRNA-like adaptors could have arisen in an RNA world. This development would have made the matching of "mRNA" sequences to amino acids more efficient, and it perhaps allowed an increase in the number of amino acids that could be used in templated protein synthesis.

Finally, the efficiency of early forms of protein synthesis would be increased dramatically by the catalysis of peptide bond formation. This evolutionary development presents no conceptual problem since, as we have seen, this reaction is catalyzed by rRNA in present-day cells. One can envision a crude 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 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.

All Present-day Cells Use DNA as Their Hereditary Material

The cells of the RNA world would presumably have been much less complex and less efficient in reproducing themselves than even the simplest present-day cells, since catalysis by RNA molecules is less efficient than that by proteins. They would have consisted of little more than a simple membrane enclosing a set of self-replicating molecules and a few other components required to provide the materials and energy for their replication. If the evolutionary speculations about RNA outlined above are correct, these early cells would also have differed fundamentally from the cells we know today in having their hereditary information stored in RNA rather than in DNA (Figure 6–101).

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.

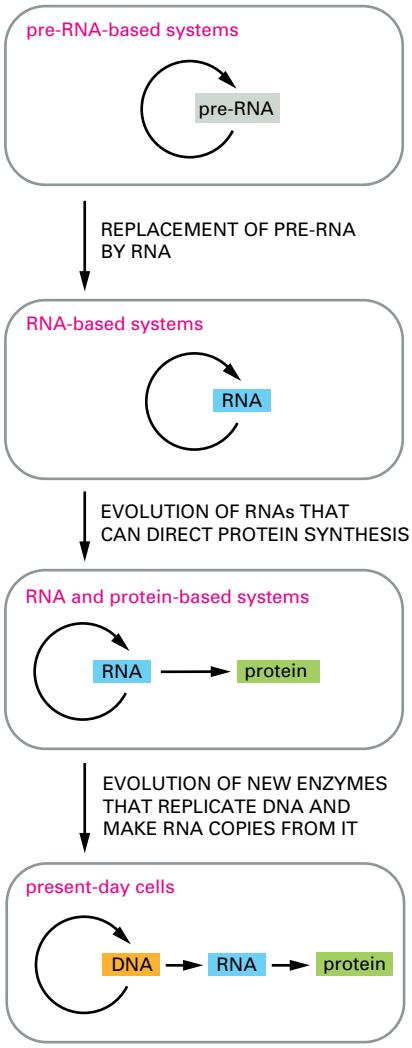


Figure 6–101 The hypothesis that RNA preceded DNA and proteins in evolution. In the earliest cells, pre-RNA molecules would have had combined genetic, structural, and catalytic functions and these functions would have gradually been replaced by RNA. In present-day cells, DNA is the repository of genetic 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 number of crucial reactions.

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. 269–272).

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 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.

References

General

Gesteland RF, Cech TR & Atkins JF (eds) (1999) The RNA World, 2nd edn. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Hartwell L, Hood L, Goldberg ML et al. (2000) Genetics: from Genes to Genomes. Boston: McGraw Hill.

Lewin B (2000) Genes VII. Oxford: Oxford University Press.

Lodish H, Berk A, Zipursky SL et al. (2000) Molecular Cell Biology, 4th edn. New York: WH Freeman.

Stent GS (1971) Molecular Genetics: An Introductory Narrative. San Francisco: WH Freeman.

Stryer L (1995) Biochemistry, 4th edn. New York: WH Freeman.

Watson JD, Hopkins NH, Roberts JW et al. (1987) Molecular Biology of the Gene, 4th edn. Menlo Park, CA: Benjamin/Cummings.

From DNA to RNA

Berget SM, Moore C & Sharp PA (1977) Spliced segments at the 5′ terminus of adenovirus 2 late mRNA. *Proc. Natl. Acad. Sci. USA* 74, 3171–3175.

Black DL (2000) Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology. *Cell* 103, 367–370.

Brenner S, Jacob F & Meselson M (1961) An unstable intermediate carrying information from genes to ribosomes for protein synthesis. *Nature* 190, 576–581.

Cech TR (1990) Nobel lecture. Self-splicing and enzymatic activity of an intervening sequence RNA from Tetrahymena. *Biosci. Rep.* 10, 239–261.

Chow LT, Gelinas RE, Broker TR et al. (1977) An amazing sequence arrangement at the 5′ ends of adenovirus 2 messenger RNA. *Cell* 12, 1–8.

Conaway JW, Shilatifard A, Dvir A & Conaway RC (2000) Control of elongation by RNA polymerase II. *Trends Biochem. Sci.* 25, 375–380.

Cramer P, Bushnell DA, Fu J et al. (2000) Architecture of RNA polymerase II and implications for the transcription mechanism. *Science* 288, 640–649.

Crick F (1979) Split genes and RNA splicing. *Science* 204, 264–271.

Daneholt B (1997) A look at messenger RNP moving through the nuclear pore. *Cell* 88, 585–588.

Darnell JE, Jr. (1985) RNA. *Sci. Am.* 253(4), 68–78.

Dvir A, Conaway JW & Conaway RC (2001) Mechanism of transcription initiation and promoter escape by RNA polymerase II. *Curr. Opin. Genet. Dev.* 11, 209–214.

Ebright RH (2000) RNA polymerase: structural similarities between bacterial RNA polymerase and eukaryotic RNA polymerase II. *J. Mol. Biol.* 304, 687–698.

Eddy SR (1999) Noncoding RNA genes. *Curr. Opin. Genet. Dev.* 9, 695–699.

Green MR (2000) TBP-associated factors (TAFIIs): multiple, selective transcriptional mediators in common complexes. *Trends Biochem. Sci.* 25, 59–63.

Harley CB & Reynolds RP (1987) Analysis of E. coli promoter sequences. *Nucleic Acids Res.* 15, 2343–2361.

Hirose Y & Manley JL (2000) RNA polymerase II and the integration of nuclear events. *Genes Dev.* 14, 1415–1429.

Kadonaga JT (1998) Eukaryotic transcription: an interlaced network of transcription factors and chromatin-modifying machines. *Cell* 92, 307–313.

Lewis JD & Tollervey D (2000) Like attracts like: getting RNA processing together in the nucleus. *Science* 288, 1385–1389.

Lisser S & Margalit H (1993) Compilation of E. coli mRNA promoter sequences. *Nucleic Acids Res.* 21, 1507–1516.

Minvielle-Sebastia L & Keller W (1999) mRNA polyadenylation and its coupling to other RNA processing reactions and to transcription. *Curr. Opin. Cell Biol.* 11, 352–357.

Mooney RA & Landick R (1999) RNA polymerase unveiled. *Cell* 98, 687–690.

Olson MO, Dundr M & Szebeni A (2000) The nucleolus: an old factory with unexpected capabilities. *Trends Cell Biol.* 10, 189–196.

Proudfoot N (2000) Connecting transcription to messenger RNA processing. *Trends Biochem. Sci.* 25, 290–293.

Reed R (2000) Mechanisms of fidelity in pre-mRNA splicing. *Curr. Opin. Cell Biol.* 12, 340–345.

Roeder RG (1996) The role of general initiation factors in transcription by RNA polymerase II. *Trends Biochem. Sci.* 21, 327–335.

Shatkin AJ & Manley JL (2000) The ends of the affair: capping and polyadenylation. *Nat. Struct. Biol.* 7, 838–842.

Smith CW & Valcarcel J (2000) Alternative pre-mRNA splicing: the logic of combinatorial control. *Trends Biochem. Sci.* 25, 381–388.

Staley JP & Guthrie C (1998) Mechanical devices of the spliceosome: motors, clocks, springs, and things. *Cell* 92, 315–326.

Tarn WY & Steitz JA (1997) Pre-mRNA splicing: the discovery of a new spliceosome doubles the challenge. *Trends Biochem. Sci.* 22, 132–137.

von Hippel PH (1998) An integrated model of the transcription complex in elongation, termination, and editing. *Science* 281, 660–665.

From RNA to Protein

Abelson J, Trotta CR & Li H (1998) tRNA splicing. *J. Biol. Chem.* 273, 12685–12688.

Anfinsen CB (1973) Principles that govern the folding of protein chains. *Science* 181, 223–230.

Cohen FE (1999) Protein misfolding and prion diseases. *J. Mol. Biol.* 293, 313–320.

Crick FHC (1966) The genetic code: III. *Sci. Am.* 215(4), 55–62.

Fedorov AN & Baldwin TO (1997) Cotranslational protein folding. *J. Biol. Chem*. 272, 32715–32718.

Frank J (2000) The ribosome—a macromolecular machine par excellence. *Chem. Biol.* 7, R133–141.

Green R (2000) Ribosomal translocation: EF-G turns the crank. *Curr. Biol.* 10, R369–373.

Hartl FU (1996) Molecular chaperones in cellular protein folding. *Nature* 381, 571–580.

Hershko A, Ciechanover A & Varshavsky A (2000) The ubiquitin system. *Nat. Med.* 6, 1073–1081.

Ibba M & Soll D (2000) Aminoacyl-tRNA synthesis. *Annu. Rev. Biochem.* 69, 617–650.

Kozak M (1999) Initiation of translation in prokaryotes and eukaryotes. *Gene* 234, 187–208.

Nissen P, Hansen J, Ban N et al. (2000) The structural basis of ribosome activity in peptide bond synthesis. *Science* 289, 920–930.

Nureki O, Vassylyev DG, Tateno M et al. (1998) Enzyme structure with two catalytic sites for double-sieve selection of substrate. *Science* 280, 578–582.

Prusiner SB (1998) Nobel lecture. Prions. *Proc. Natl. Acad. Sci. USA* 95, 13363–13383.

Rich A & Kim SH (1978) The three-dimensional structure of transfer RNA. *Sci. Am.* 238(1), 52–62.

Sachs AB & Varani G (2000) Eukaryotic translation initiation: there are (at least) two side to every story. *Nat. Struct. Biol.* 7, 356–361.

The Genetic Code. (1966) Cold Spring Harbor Symposium on Quantitative Biology, vol XXXI. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Turner GC & Varshavsky A (2000) Detecting and measuring cotranslational protein degradation *in vivo*. *Science* 289, 2117–2120.

Varshavsky A, Turner G, Du F et al. (2000) The ubiquitin system and the N-end rule pathway. *Biol. Chem.* 381, 779–789.

Voges D, Zwickl P & Baumeister W (1999) The 26S proteasome: a molecular machine designed for controlled proteolysis. *Annu. Rev. Biochem.* 68, 1015–1068.

Wilson KS & Noller HF (1998) Molecular movement inside the translational engine. *Cell* 92, 337–349.

Wimberly BT, Brodersen DE, Clemons WM et al. (2000) Structure of the 30S ribosomal subunit. *Nature* 407, 327–339.

Yusupov MM, Yusupova GZ, Baucom A et al. (2001) Crystal structure of the ribosome at 5.5A resolution. *Science* 292, 883–896.

The RNA World and the Origins of Life

Bartel DP & Unrau PJ (1999) Constructing an RNA world. *Trends Cell Biol.* 9, M9–M13.

Joyce GF (1992) Directed molecular evolution. *Sci. Am.* 267(6), 90–97.

Knight RD & Landweber LF (2000) The early evolution of the genetic code. *Cell* 101, 569–572.

Orgel L (2000) Origin of life. A simpler nucleic acid. *Science* 290, 1306–1307.

Szathmary E (1999) The origin of the genetic code: amino acids as cofactors in an RNA world. *Trends Genet.* 15, 223–229.