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

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

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

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

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

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

Table 6–1 Principal Types of RNAs Produced in Cells

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

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

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

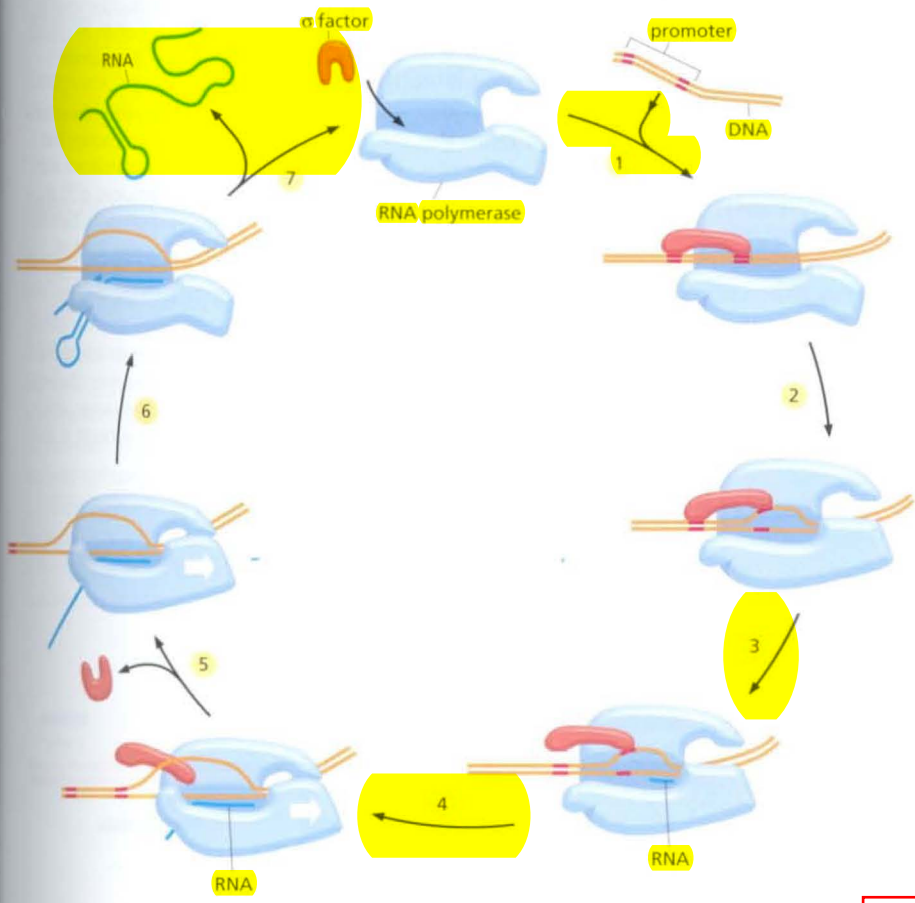
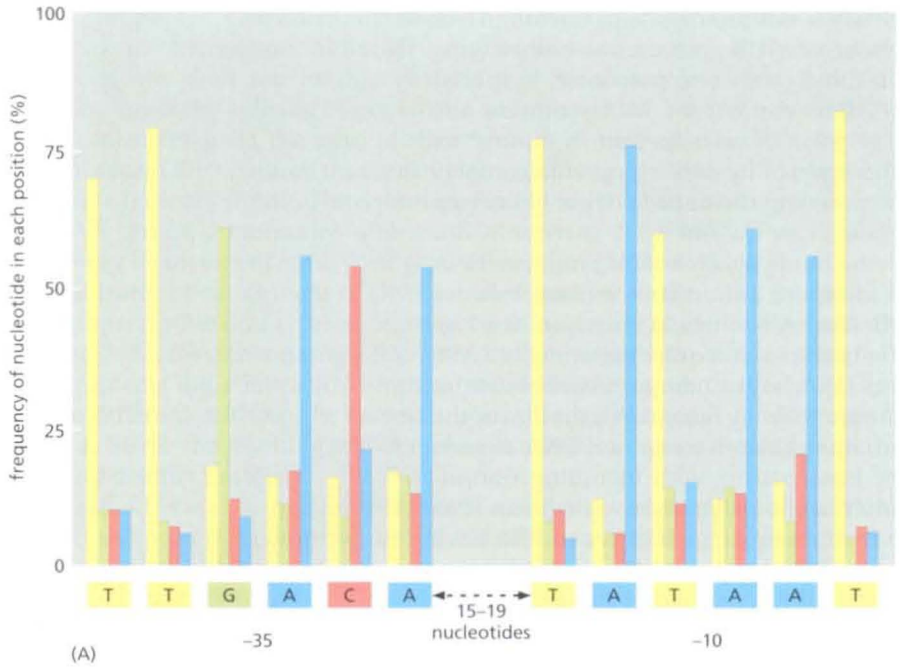



**Figure 6–11 The transcription cycle of bacterial RNA polymerase.** In step 1, the RNA polymerase holoenzyme (polymerase core enzyme plus σ factor) assembles and then locates a promoter (see Figure 6–12). The polymerase unwinds the DNA at the position at which transcription is to begin (step 2) and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

2nd



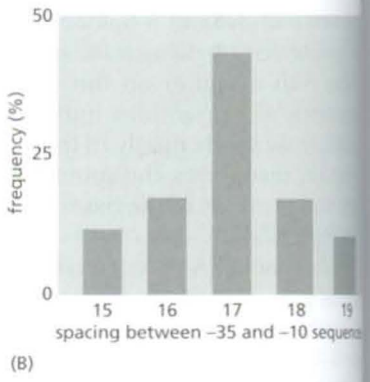




**Figure 6–12 Consensus sequence for the major class of *E. coli* promoters.** (A) The promoters are characterized by two hexameric DNA sequences, the –35 sequence and the –10 sequence named for their approximate location relative to the start point of transcription (designated +1). For convenience, 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 $\sigma^{70}$). 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 $\sigma^{70}$ have an additional sequence, located upstream (to the *left*, in the figure) of the –35 hexamer, which is recognized by another subunit of RNA polymerase.

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

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

## Transcription Start and Stop Signals Are Heterogeneous in Nucleotide Sequence

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

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

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

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

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

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

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

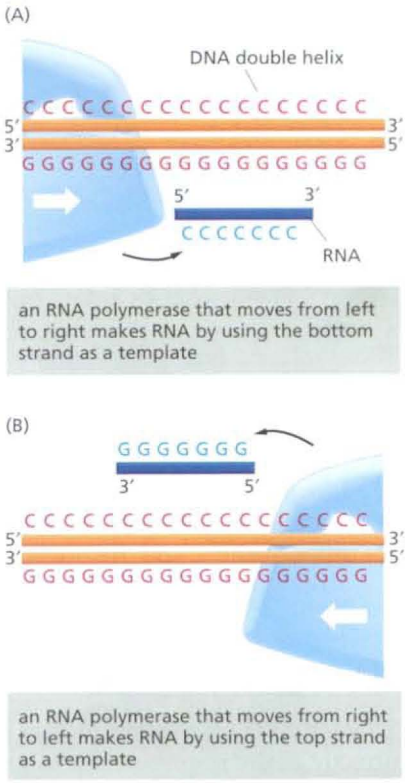



**Figure 6–13 The importance of RNA polymerase orientation.** The DNA strand serving as template must be traversed in a 3′-to-5′ direction. Thus, the direction of RNA polymerase movement determines which of the two DNA strands is to serve as a template for the synthesis of RNA, as shown in (A) and (B).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

## Transcription Initiation in Eucaryotes Requires Many Proteins

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

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

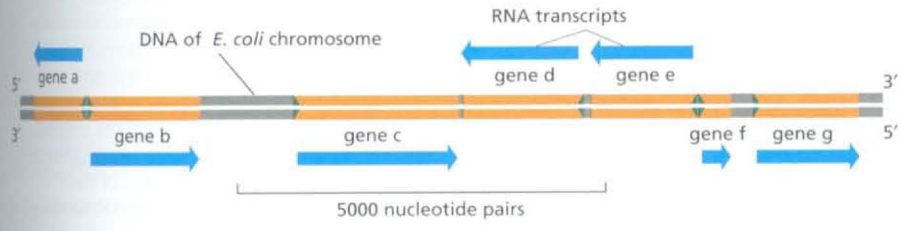



**Figure 6–14 Directions of transcription along a short portion of a bacterial chromosome.** Some genes are transcribed using one DNA strand as a template, while others are transcribed using the other DNA strand. The direction of transcription is determined by the promoter at the beginning of each gene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

Table 6–2 The Three RNA Polymerases in Eucaryotic Cells

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

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

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

## RNA Polymerase II Requires General Transcription Factors

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

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

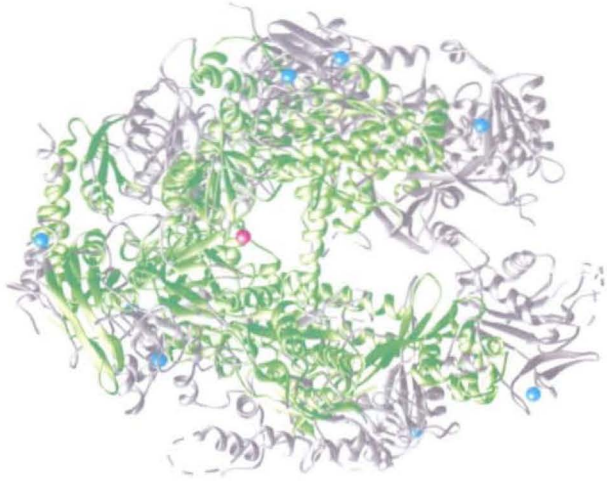


**Figure 6–15 Structural similarity between a bacterial RNA polymerase and a eucaryotic RNA polymerase II.** Regions of the two RNA polymerases that have similar structures are indicated in *green*. The eucaryotic polymerase is larger than the bacterial enzyme (12 subunits instead of 5), and some of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

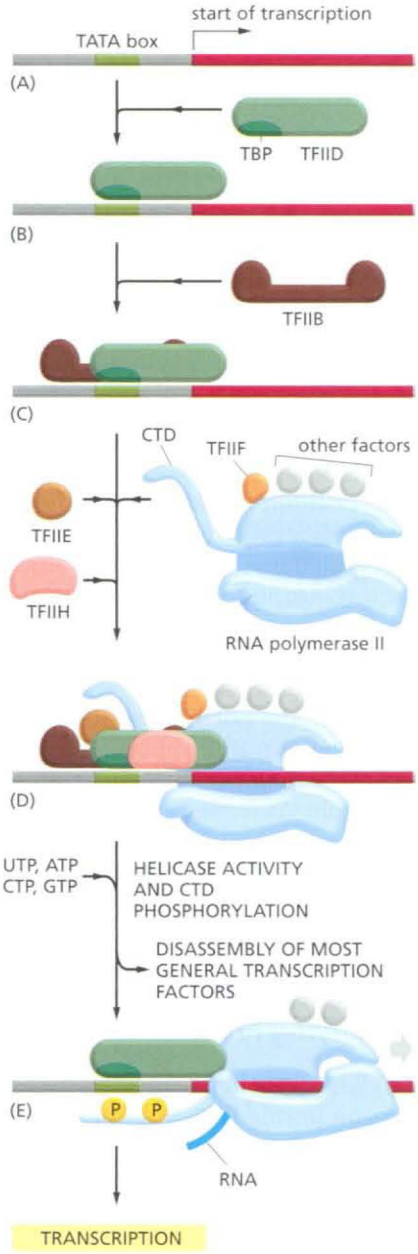



**Figure 6–16 Initiation of transcription of a eucaryotic gene by RNA polymerase II.** To begin transcription, RNA polymerase requires several general transcription factors. (A) The promoter contains a DNA sequence called the TATA box, which is located 25 nucleotides away from the site at which transcription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

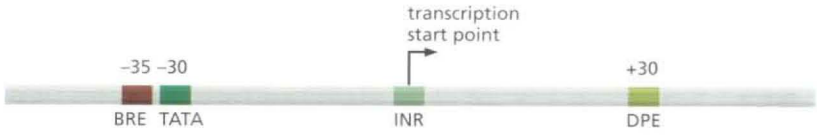
of the TATA box (**Figure 6–18**). This distortion is thought to serve as a physical landmark for the location of an active promoter in the midst of a very large genome, and it brings DNA sequences on both sides of the distortion together to allow for subsequent protein assembly steps. Other factors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

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

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

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

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



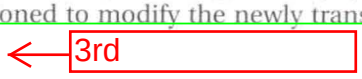

| element | consensus sequence | general transcription factor |
|---|---|---|
| BRE | G/C G/C G/A C G C C | TFIIB |
| TATA | T A T A A/T A A/T | TBP |
| INR | C/T C/T A N T/A C/T C/T | TFIID |
| DPE | A/G G A/T C G T G | TFIID |

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

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

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

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




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

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

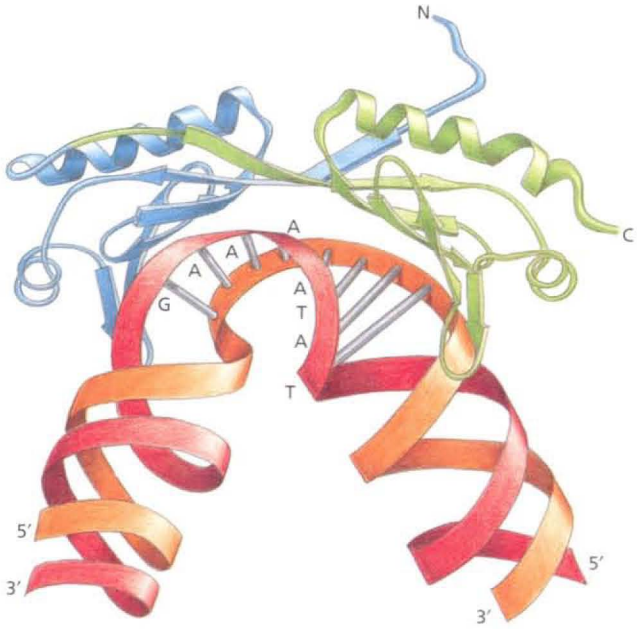



**Figure 6–18 Three-dimensional structure of TBP (TATA-binding protein) bound to DNA.** The TBP is the subunit of the general transcription factor TFIID that is responsible for recognizing and binding to the TATA box sequence in the DNA *(red)*. The unique DNA bending caused by TBP—two kinks in the double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

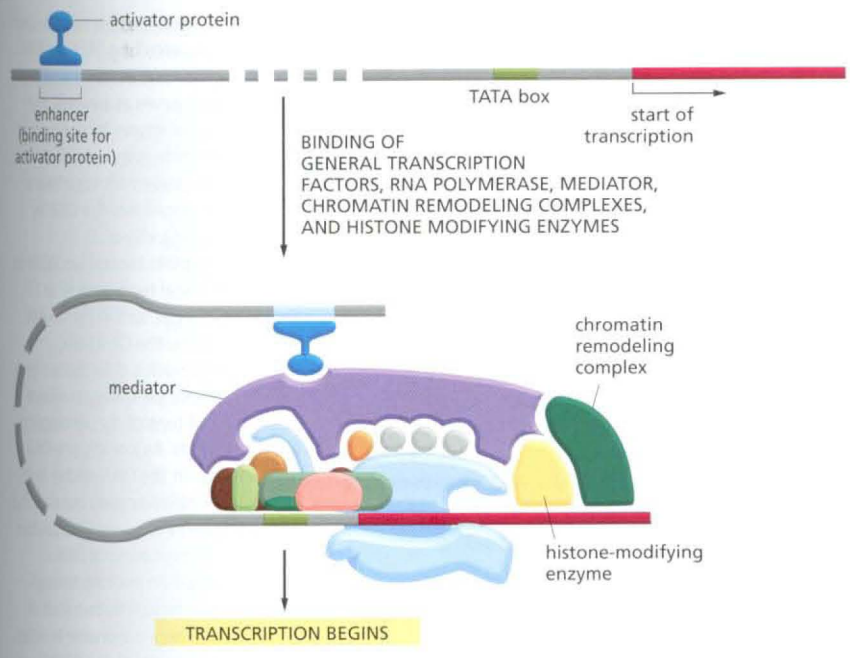



**Figure 6–19 Transcription initiation by RNA polymerase II in a eucaryotic cell.** Transcription initiation *in vivo* requires the presence of transcriptional activator proteins. As described in Chapter 7, these proteins bind to specific short sequences in DNA. Although only one is shown here, a typical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

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

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

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

## Transcription Elongation Produces Superhelical Tension in DNA

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

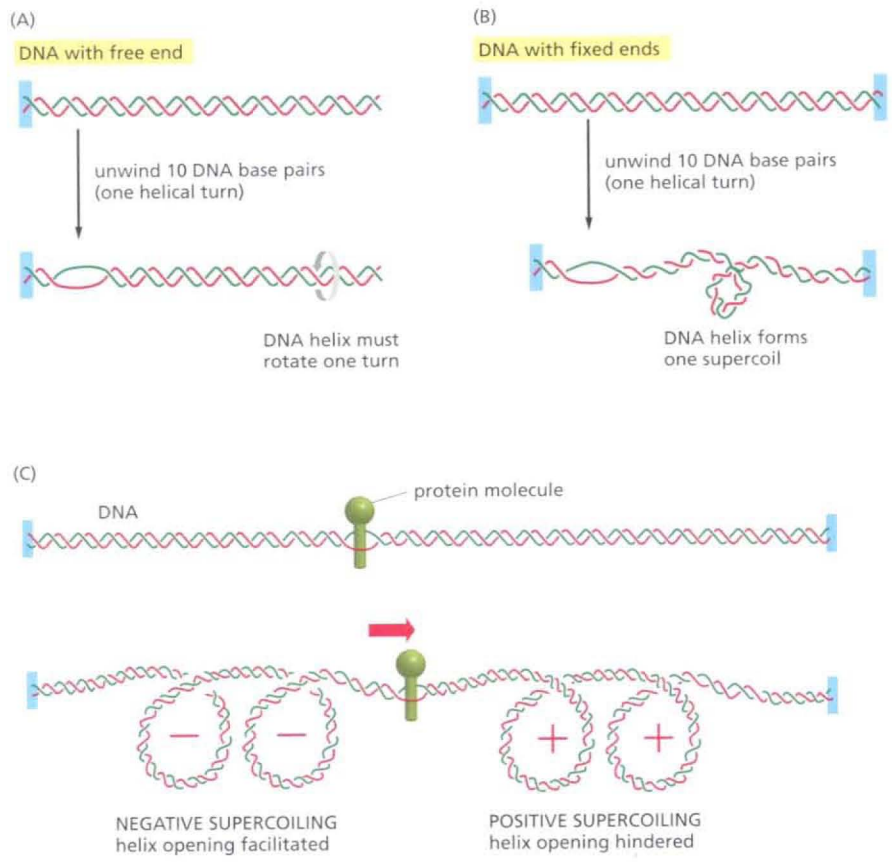



**Figure 6–20 Superhelical tension in DNA causes DNA supercoiling.** (A) For a DNA molecule with one free end (or a nick in one strand that serves as a swivel), the DNA double helix rotates by one turn for every 10 nucleotide pairs opened. (B) If rotation is prevented, superhelical tension is introduced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

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

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

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

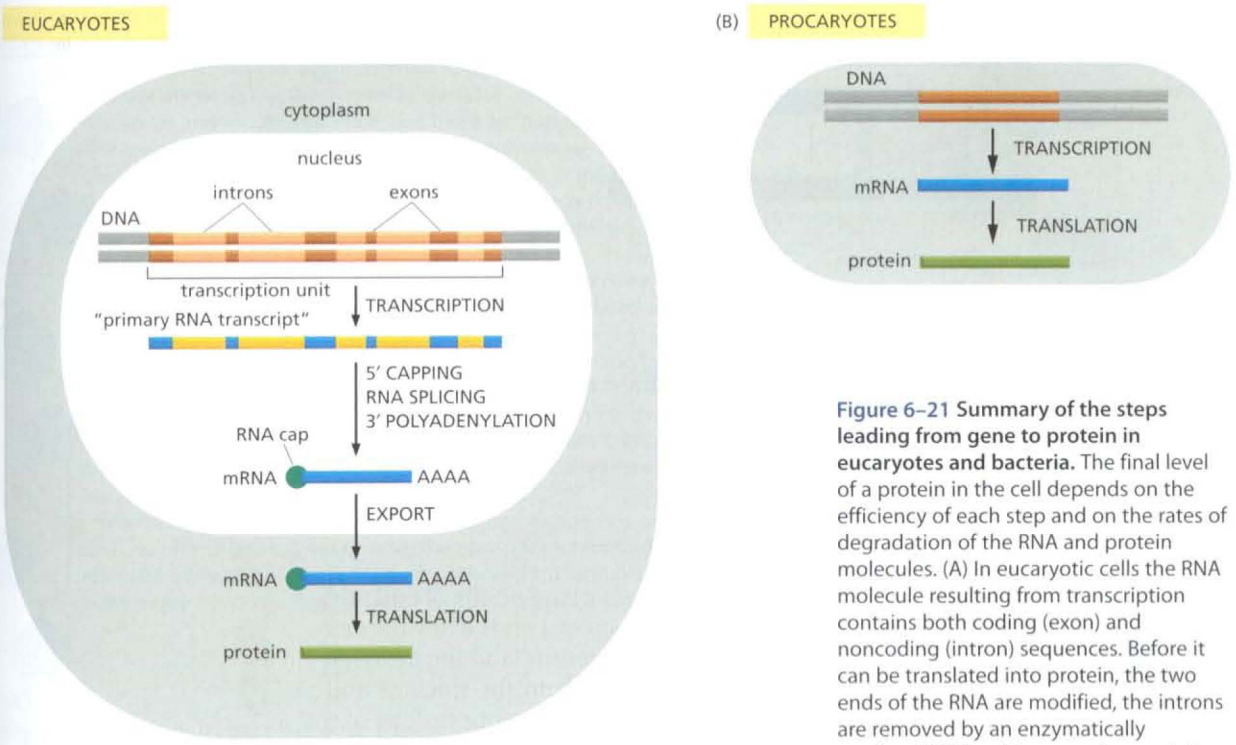



**Figure 6–21 Summary of the steps leading from gene to protein in eucaryotes and bacteria.** The final level of a protein in the cell depends on the efficiency of each step and on the rates of degradation of the RNA and protein molecules. (A) In eucaryotic cells the RNA molecule resulting from transcription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

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

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

## Transcription Elongation in Eucaryotes Is Tightly Coupled to RNA Processing

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

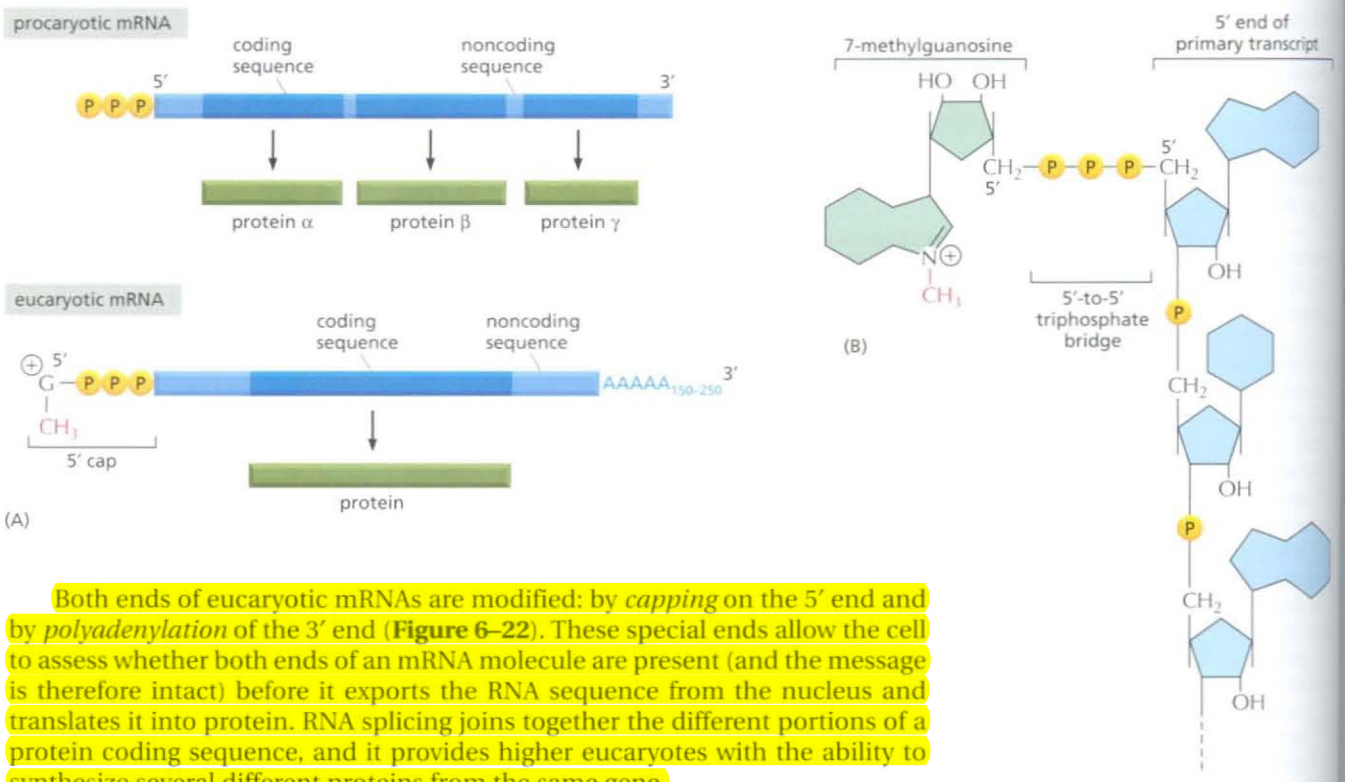



Figure 6–22 **A comparison of the structures of procaryotic and eucaryotic mRNA molecules.** (A) The 5′ and 3′ ends of a bacterial mRNA are the unmodified ends of the chain synthesized by the RNA polymerase, which initiates and terminates transcription at those points, respectively. The corresponding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

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

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

## RNA Capping Is the First Modification of Eucaryotic Pre-mRNAs

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

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

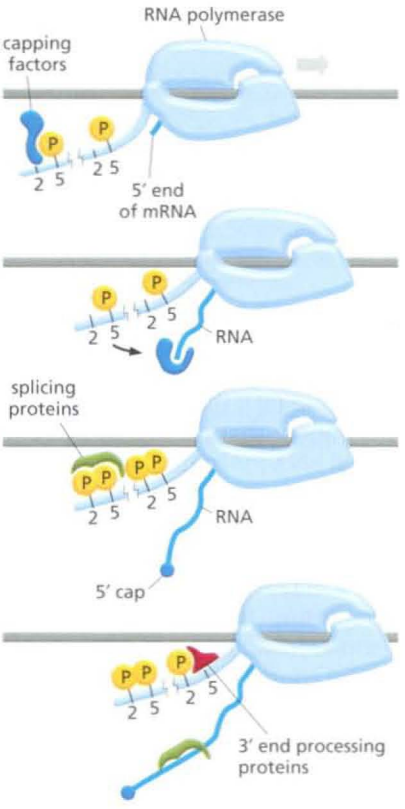



**Figure 6–23 Eucaryotic RNA polymerase II as an "RNA factory."** As the polymerase transcribes DNA into RNA, it carries pre-mRNA-processing proteins on its tail that are transferred to the nascent RNA at the appropriate time. The tail, known as the CTD, contains 52 tandem repeats of a seven amino acid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

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

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

## RNA Splicing Removes Intron Sequences from Newly Transcribed Pre-mRNAs

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

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

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

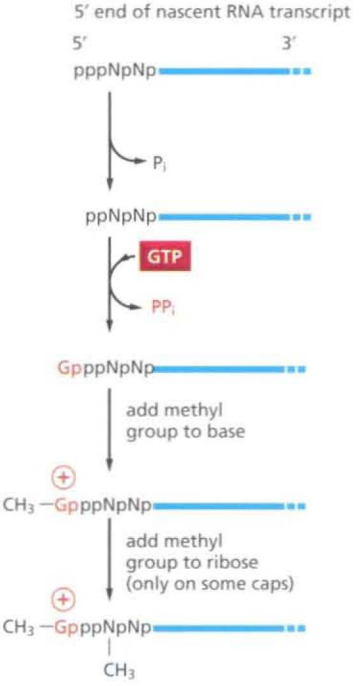



**Figure 6–24 The reactions that cap the 5′ end of each RNA molecule synthesized by RNA polymerase II.** The final cap contains a novel 5′-to-5′ linkage between the positively charged 7-methyl G residue and the 5′ end of the RNA transcript (see Figure 6–22B). The letter N represents any one of the four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

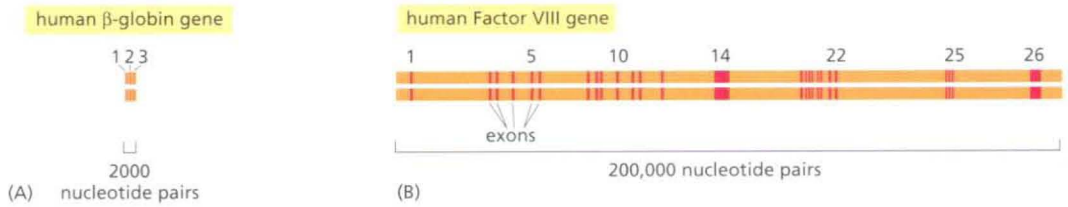



**Figure 6–25 Structure of two human genes showing the arrangement of exons and introns.** (A) The relatively small β-globin gene, which encodes one of the subunits of the oxygen-carrying protein hemoglobin, contains 3 exons (see also Figure 4–7). (B) The much larger Factor VIII gene contains 26 exons;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

principle take place without nucleoside triphosphate hydrolysis. However, the machinery that catalyzes pre-mRNA splicing is complex, consisting of 5 additional RNA molecules and as many as 200 proteins, and it hydrolyzes many ATP molecules per splicing event. This additional complexity ensures that splicing is accurate, while at the same time being flexible enough to deal with the enormous variety of introns found in a typical eucaryotic cell. 4th

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

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

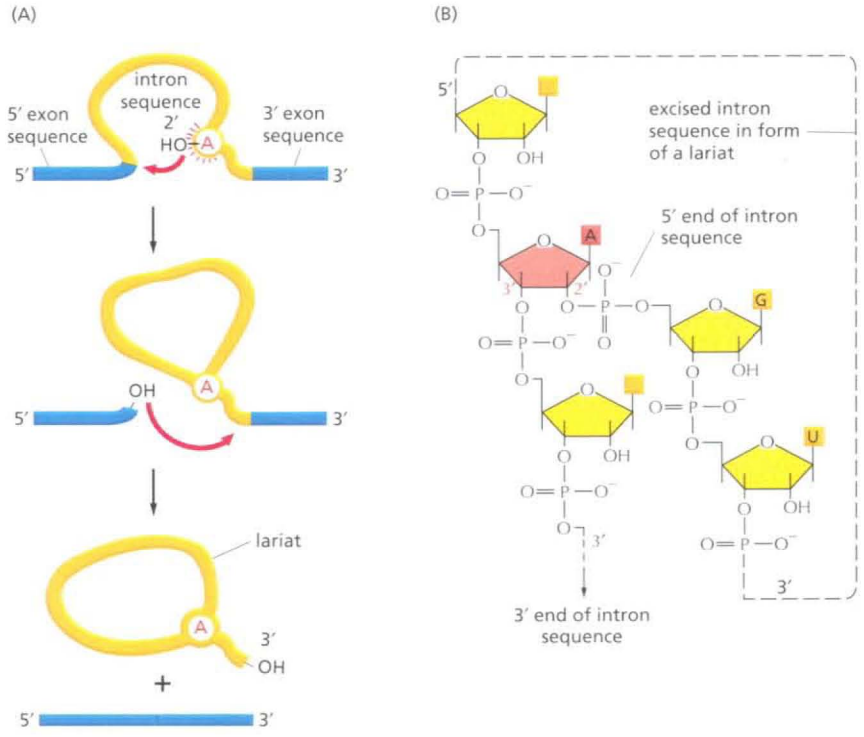



**Figure 6–26 The pre-mRNA splicing reaction.** (A) In the first step, a specific adenine nucleotide in the intron sequence (indicated in *red*) attacks the 5′ splice site and cuts the sugar-phosphate backbone of the RNA at this point. The cut 5′ end of the intron becomes covalently linked to the adenine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

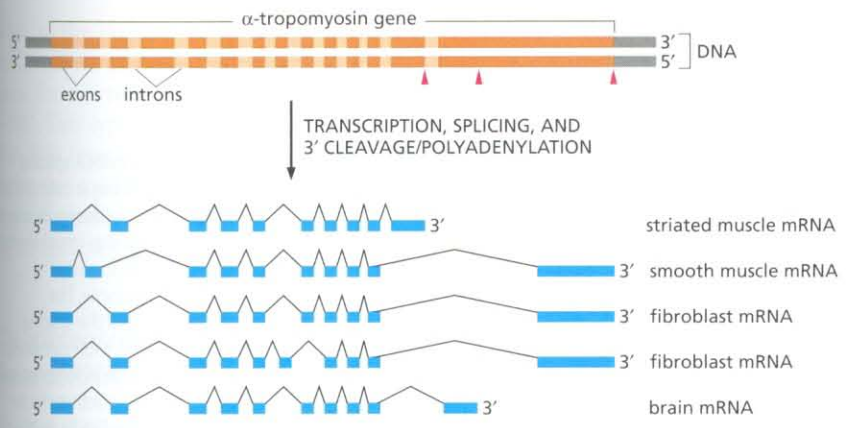



**Figure 6–27 Alternative splicing of the α-tropomyosin gene from rat.** α-Tropomyosin is a coiled-coil protein (see Figure 3–9) that regulates contraction in muscle cells. The primary transcript can be spliced in different ways, as indicated in the figure, to produce distinct mRNAs, which then give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

## Nucleotide Sequences Signal Where Splicing Occurs

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

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

## RNA Splicing Is Performed by the Spliceosome

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

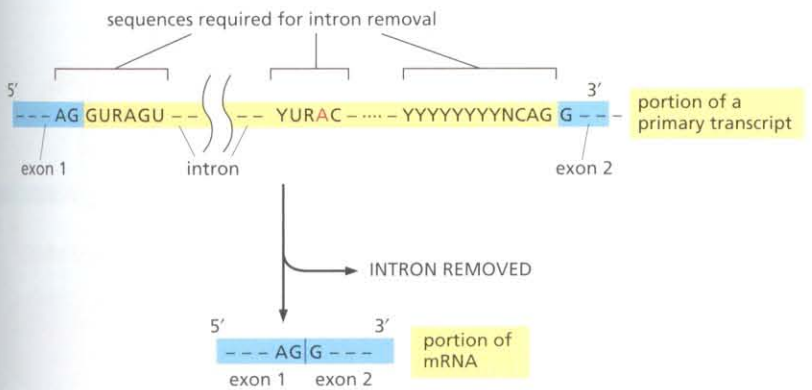
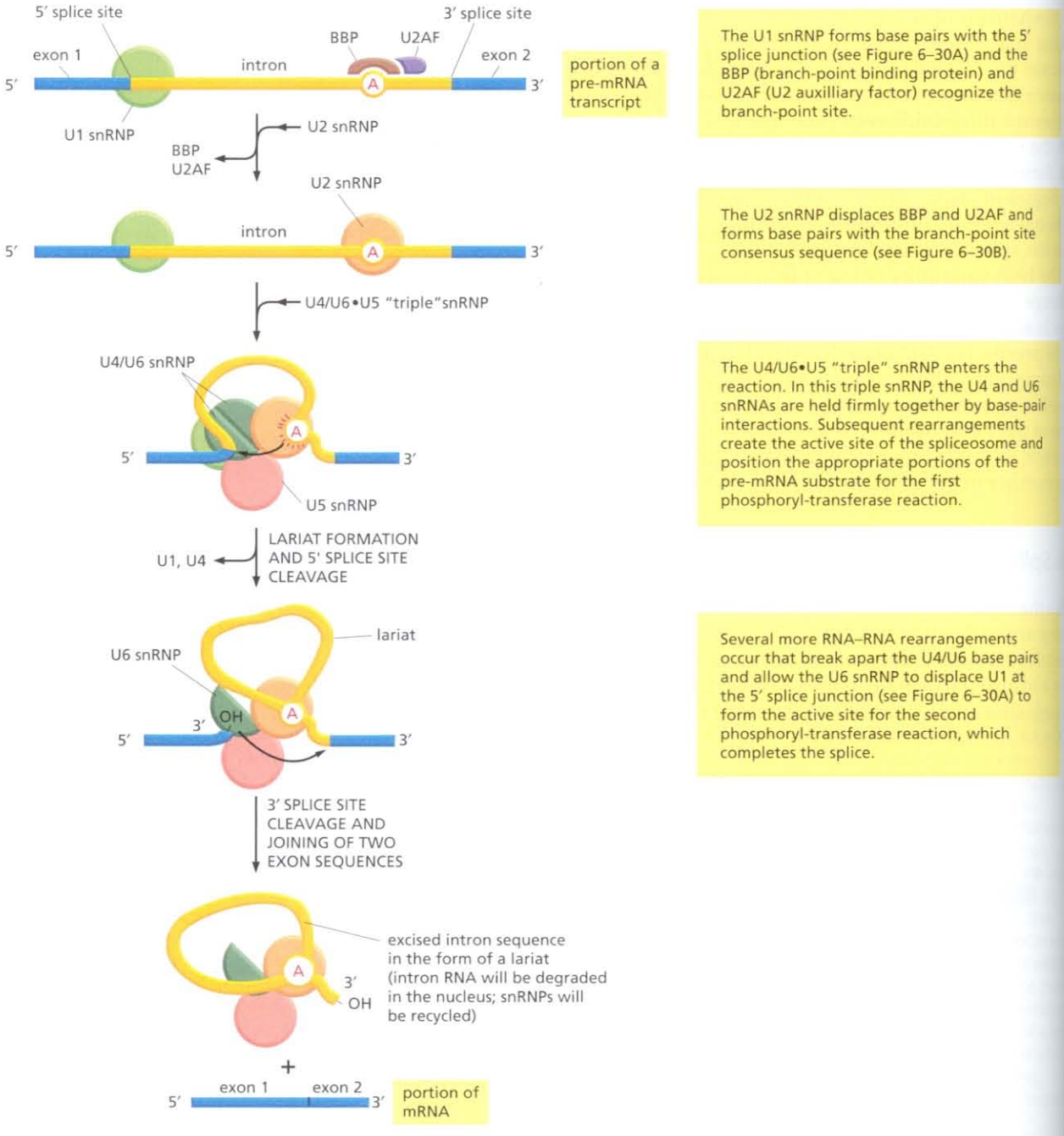



**Figure 6–28 The consensus nucleotide sequences in an RNA molecule that signal the beginning and the end of most introns in humans.** Only the three blocks of nucleotide sequences shown are required to remove an intron sequence; the rest of the intron can be occupied by any nucleotides. Here A, G, U,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

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

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




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

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

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

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

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

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

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

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

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

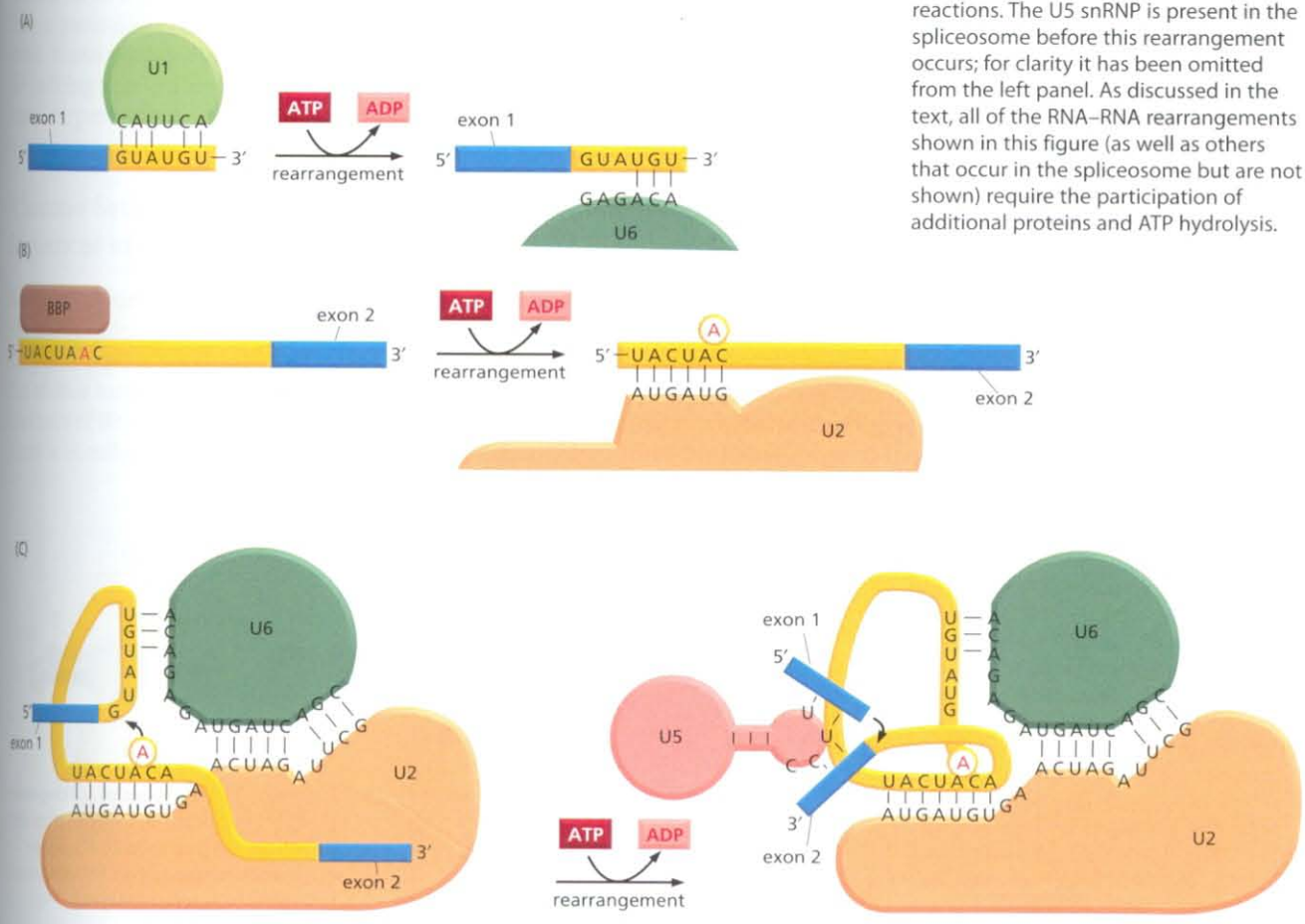



**Figure 6–30 Several of the rearrangements that take place in the spliceosome during pre-mRNA splicing.** Shown here are the details for the yeast *Saccharomyces cerevisiae*, in which the nucleotide sequences involved are slightly different from those in human cells. (A) The exchange of U1 snRNP for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

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

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

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

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

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

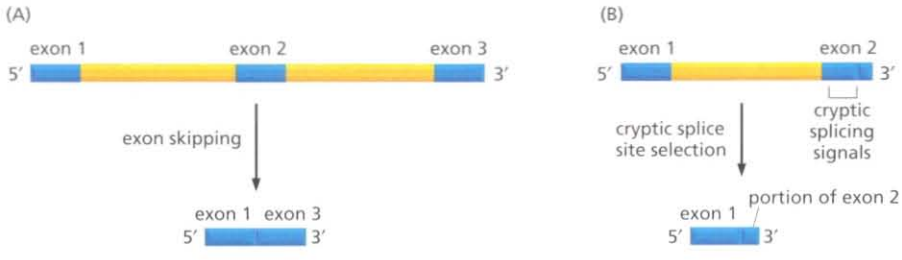
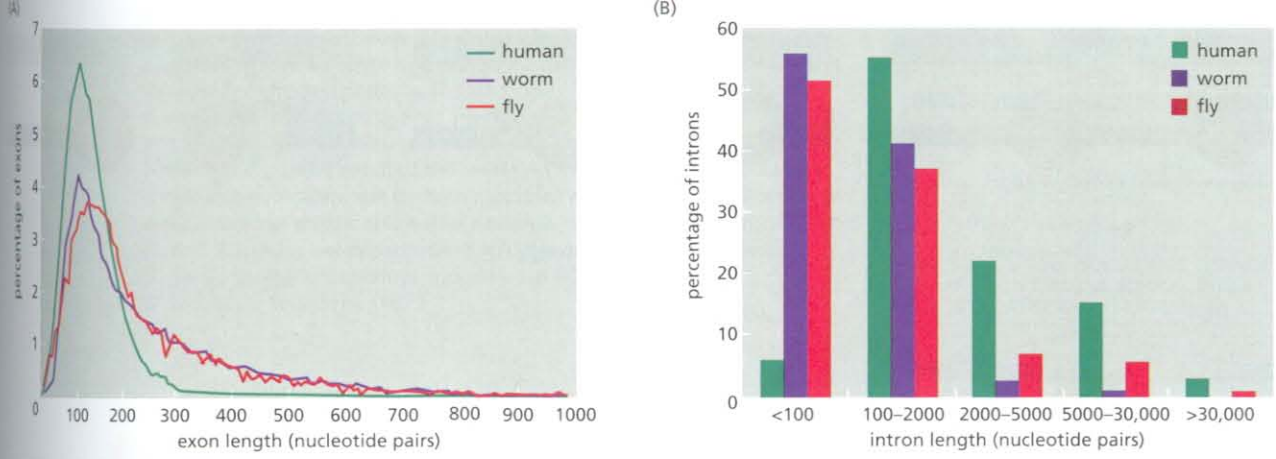



**Figure 6–31 Two types of splicing err** (A) Exon skipping. (B) Cryptic splice-site selection. Cryptic splicing signals are nucleotide sequences of RNA that clos resemble true splicing signals.




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

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

5th

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

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

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

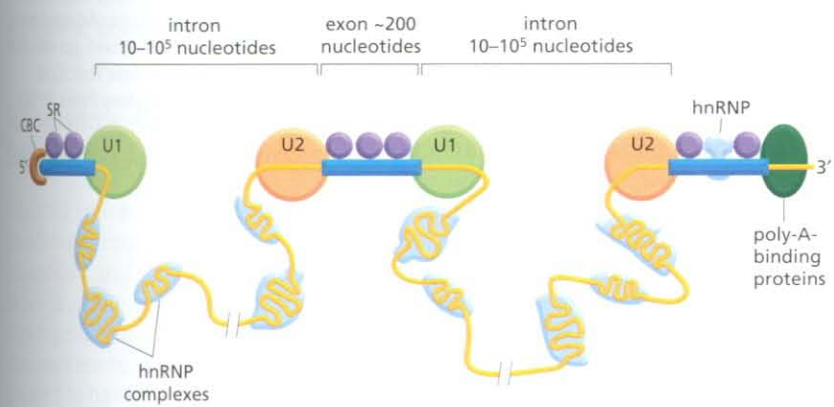
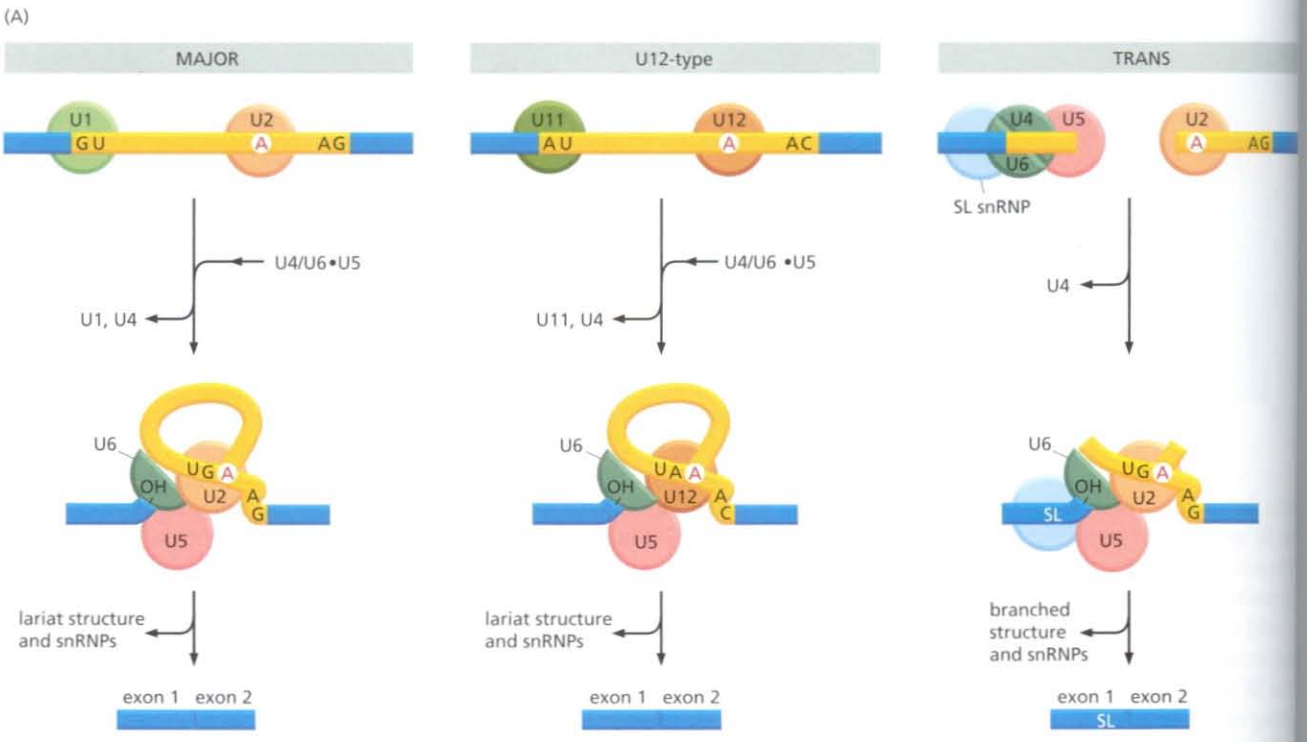



**Figure 6–33 The exon definition idea.** According to one proposal, SR proteins bind to each exon sequence in the pre-mRNA and thereby help to guide the snRNPs to the proper intron/exon boundaries. This demarcation of exons by the SR proteins occurs co-transcriptionally, beginning at the CBC (cap-binding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



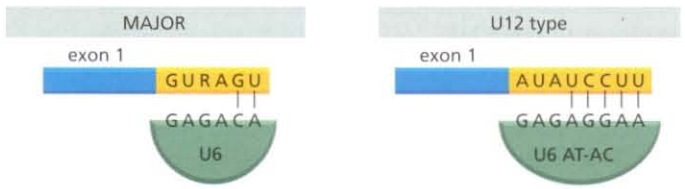




**Figure 6–34 Outline of the mechanisms used for three types of RNA splicing.** (A) Three types of spliceosomes. The major spliceosome *(left)*, the U12-type spliceosome *(middle)*, and the trans-spliceosome *(right)* are each shown at two stages of assembly. Introns removed by the U12-type spliceosome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

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

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

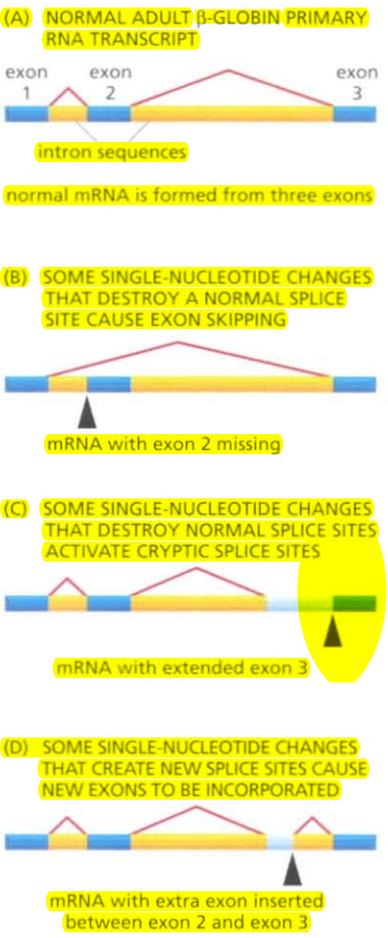



**Figure 6–35 Abnormal processing of the β-globin primary RNA transcript in humans with the disease β thalassemia.** In the examples shown, the disease is caused by splice-site mutations (*black arrowheads*) found in the genomes of affected patients. The *dark blue boxes* represent the three normal exon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

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

## RNA Splicing Shows Remarkable Plasticity

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

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

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

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

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

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

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

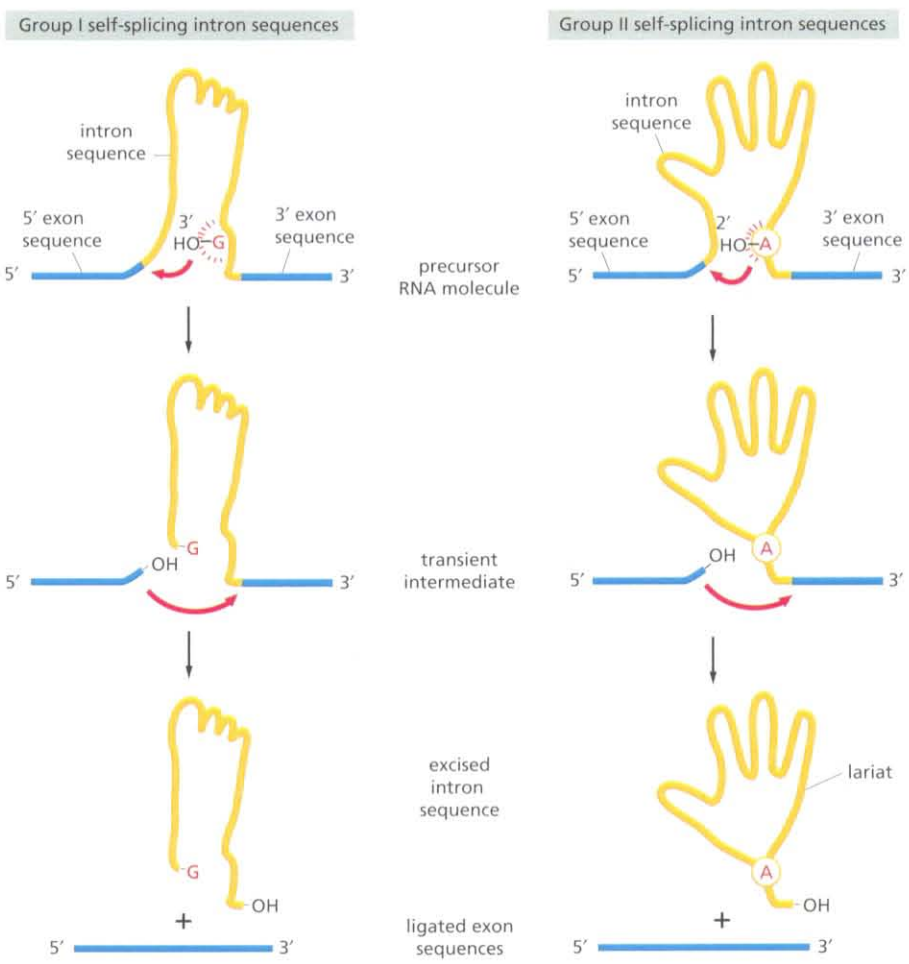



**Figure 6–36 The two known classes of self-splicing intron sequences.** The figure emphasizes the similarities between the two mechanisms. Both are normally aided by proteins in the cell that speed up the reaction, but the catalysis is nevertheless mediated by the RNA in the intron sequence. The group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

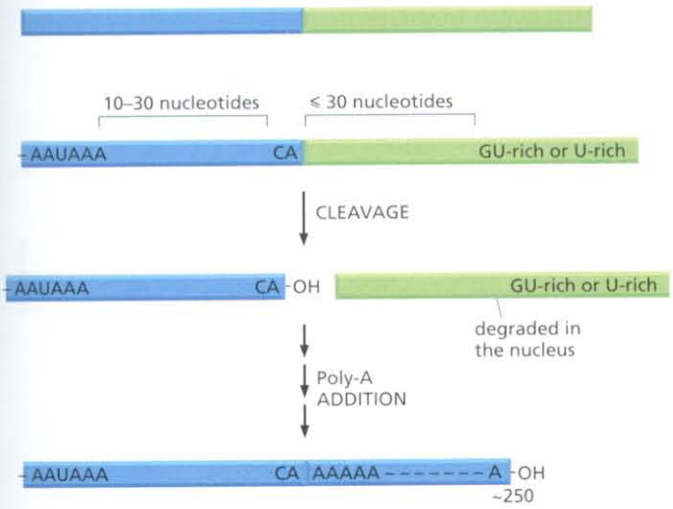



**Figure 6–37 Consensus nucleotide sequences that direct cleavage and polyadenylation to form the 3′ end of a eucaryotic mRNA.** These sequences are encoded in the genome; specific proteins recognize them after they are transcribed into RNA. The hexamer AAUAAA is bound by CPSF, the GU-rich element beyond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

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

6th

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

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

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

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

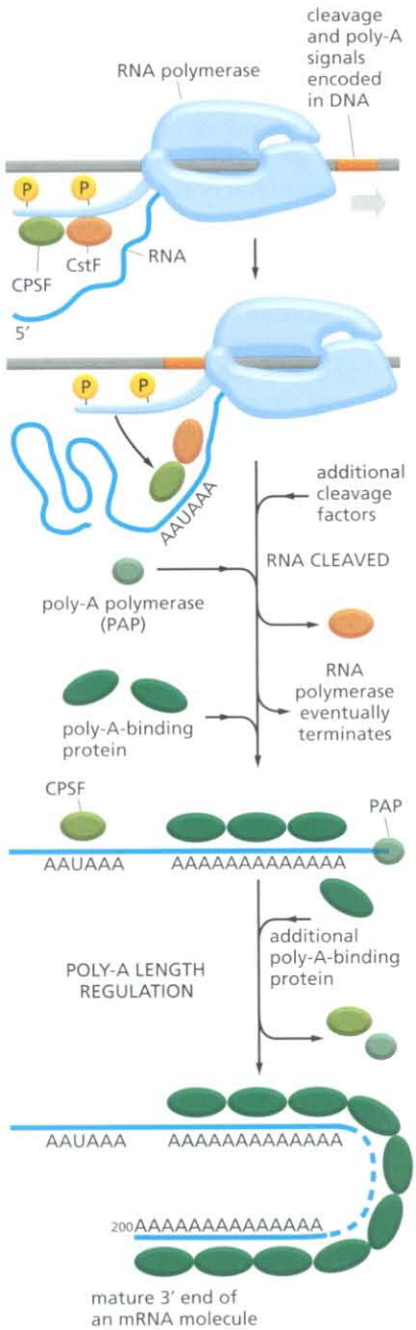



**Figure 6–38 Some of the major steps in generating the 3′ end of a eucaryotic mRNA.** This process is much more complicated than the analogous process in bacteria, where the RNA polymerase simply stops at a termination signal and releases both the 3′ end of its transcript and the DNA template (see Figure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

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

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

## Mature Eucaryotic mRNAs Are Selectively Exported from the Nucleus

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

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

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

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

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

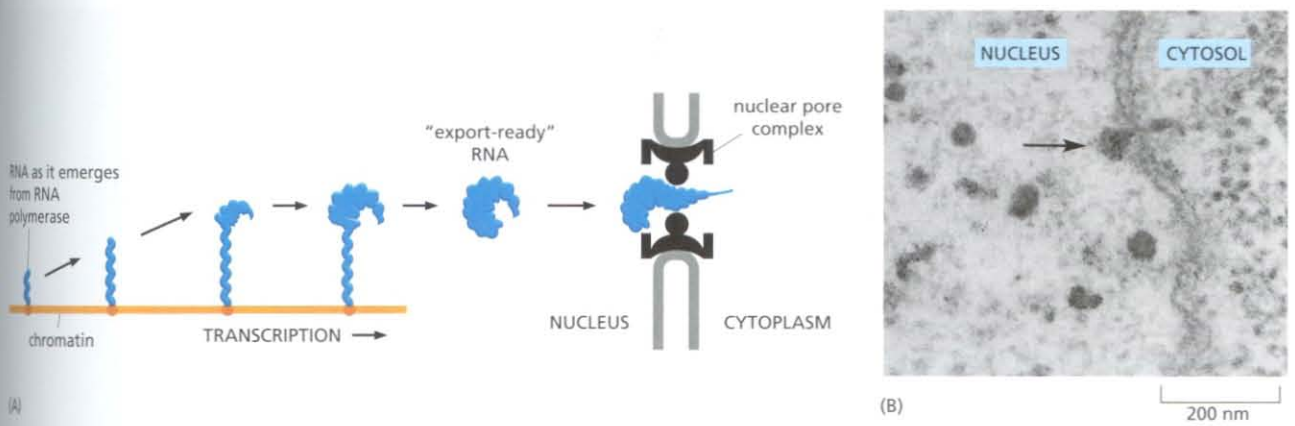



**Figure 6–39 Transport of a large mRNA molecule through the nuclear pore complex.** (A) The maturation of an mRNA molecule as it is synthesized by RNA polymerase and packaged by a variety of nuclear proteins. This drawing of an unusually abundant RNA, called the Balbiani Ring mRNA, is based on EM micrographs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

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

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

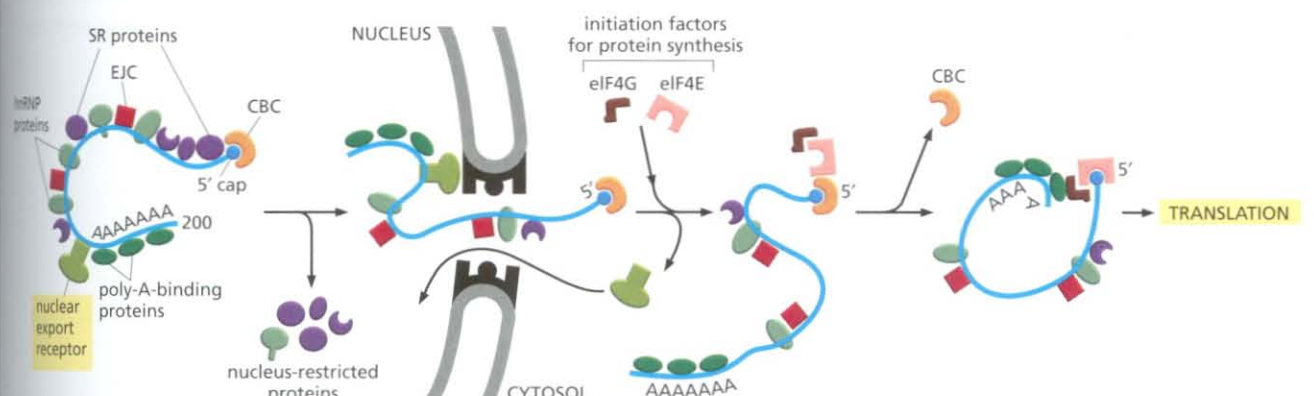



**Figure 6–40 Schematic illustration of an "export-ready" mRNA molecule and its transport through the nuclear pore.** As indicated, some proteins travel with the mRNA as it moves through the pore, whereas others remain in the nucleus. The nuclear export receptor for mRNAs is a complex of proteins that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

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

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

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

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

## Many Noncoding RNAs Are Also Synthesized and Processed in the Nucleus

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

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

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


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

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

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

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

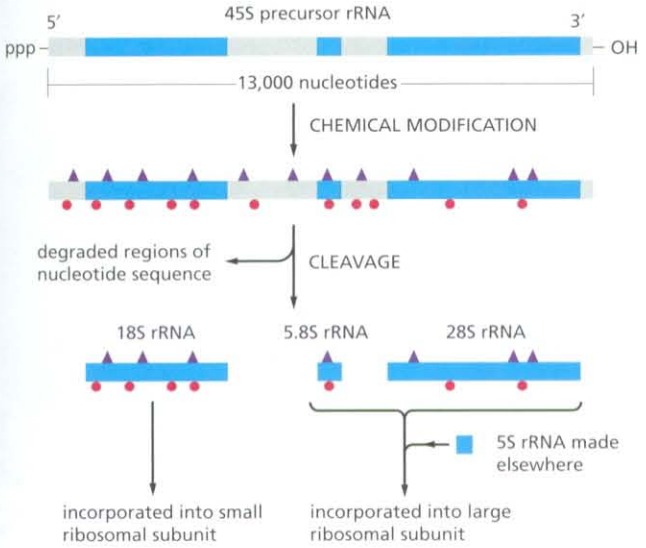



**Figure 6–42 The chemical modification and nucleolytic processing of a eucaryotic 45S precursor rRNA molecule into three separate ribosomal RNAs.** Two types of chemical modifications (color-coded as indicated in Figure 6–43) are made to the precursor rRNA before it is cleaved. Nearly half of the nucleotide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

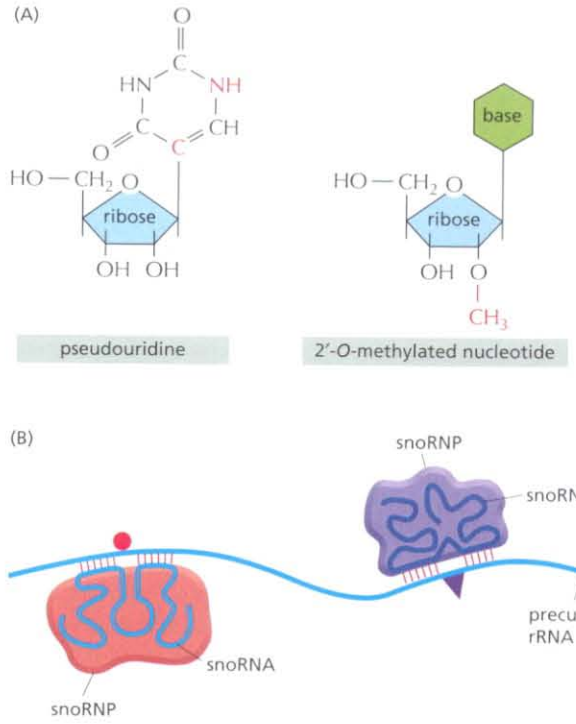



**Figure 6–43 Modifications of the precursor rRNA by guide RNAs.** (A) Two prominent covalent modifications occur after rRNA synthesis; the differences from the initially incorporated nucleotide are indicated by *red* atoms. Pseudouridine is an isomer of uridine; the base has been "rotated" relative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

## The Nucleolus Is a Ribosome-Producing Factory

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

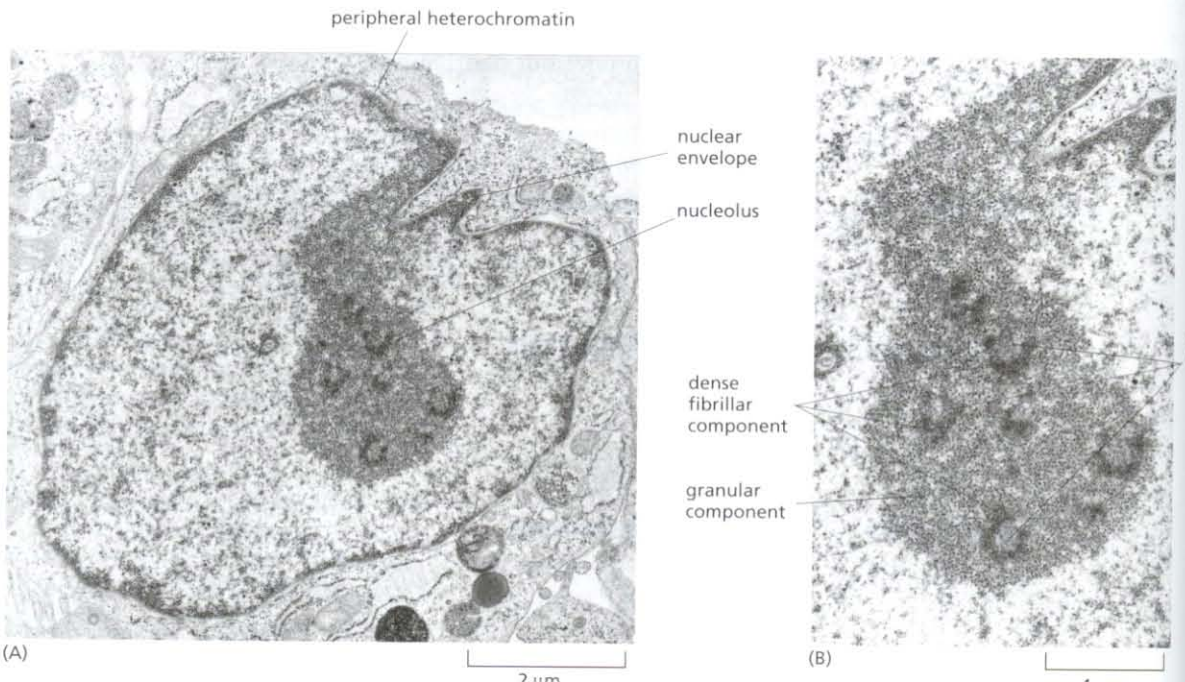



**Figure 6–44 Electron micrograph of a thin section of a nucleolus in a human fibroblast, showing its three distinct zones.** (A) View of entire nucleus. (B) High-power view of the nucleolus. It is believed that transcription of the rRNA genes takes place between the fibrillar center and the dense fibrillar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

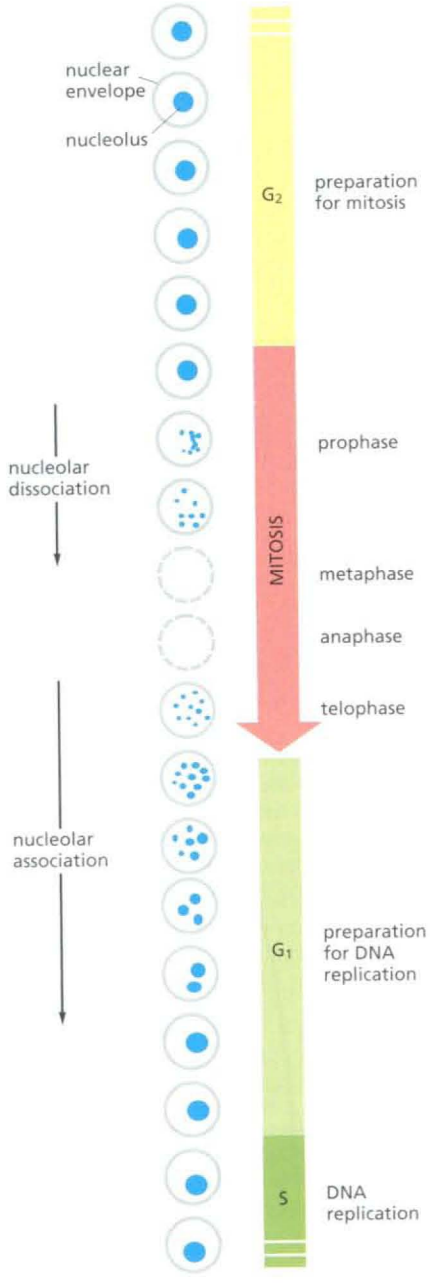



**Figure 6–45 Changes in the appearance of the nucleolus in a human cell during the cell cycle.** Only the cell nucleus is represented in this diagram. In most eucaryotic cells the nuclear envelope breaks down during mitosis, as indicated by the dashed circles.

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

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

## The Nucleus Contains a Variety of Subnuclear Structures

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

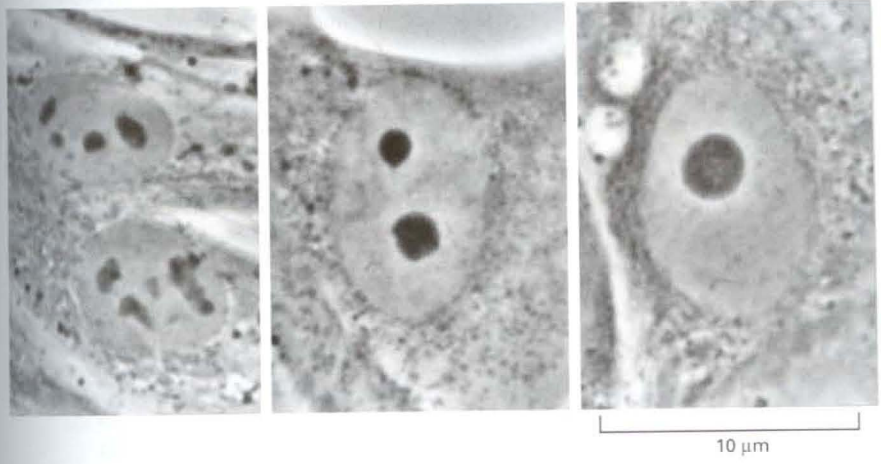



**Figure 6–46 Nucleolar fusion.** These light micrographs of human fibroblasts grown in culture show various stages of nucleolar fusion. After mitosis, each of the 10 human chromosomes that carry a cluster of rRNA genes begins to form a tiny nucleolus, but these rapidly coalesce as they grow to form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

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

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

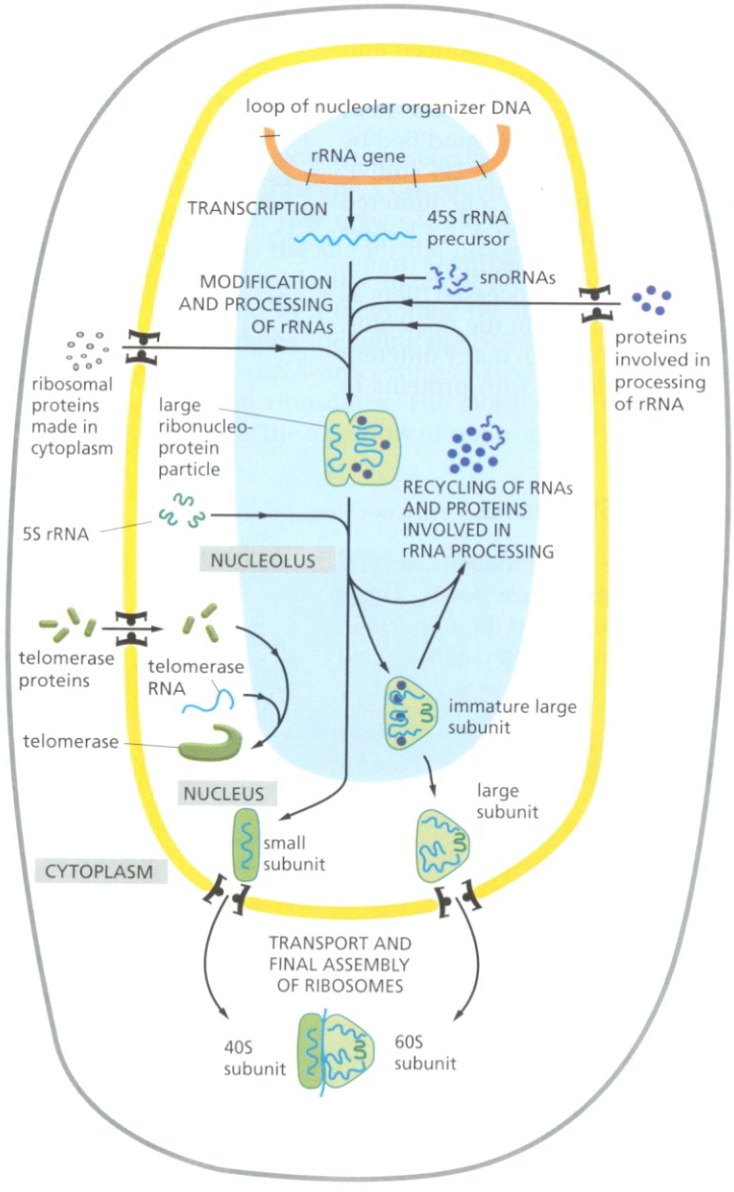



**Figure 6–47 The function of the nucleolus in ribosome and other ribonucleoprotein synthesis.** The 45S precursor rRNA is packaged in a large ribonucleoprotein particle containing many ribosomal proteins imported from the cytoplasm. While this particle remains at the nucleolus, selected pieces are added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

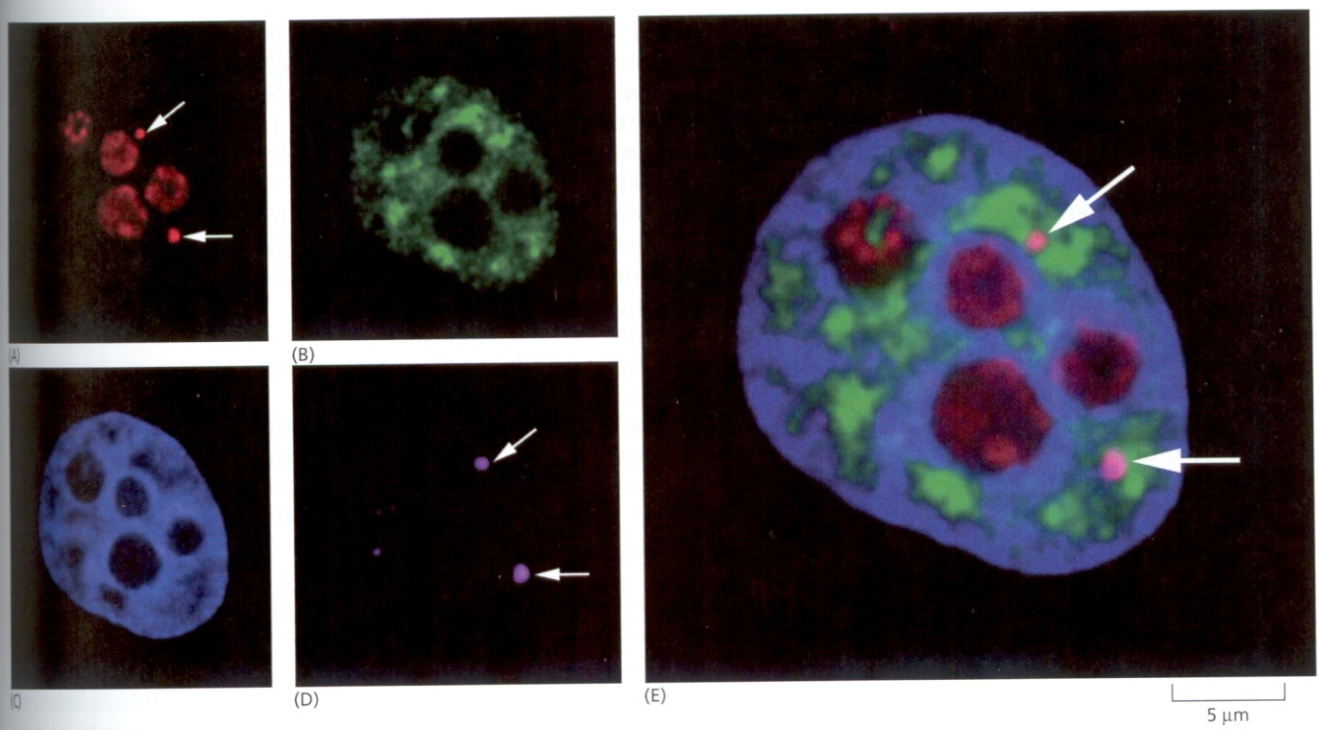



**Figure 6–48 Visualization of some prominent nuclear bodies.** (A)–(D) Micrographs of the same human cell nucleus, each processed to show a particular set of nuclear structures. (E) All four images enlarged and superimposed. (A) shows the location of the protein fibrillarin (a component of several snoRNPs),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

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

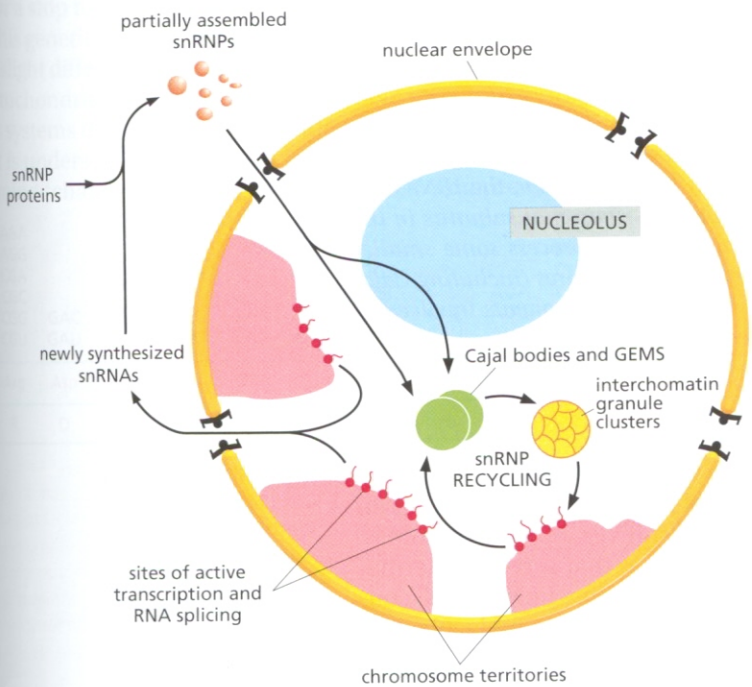



**Figure 6–49 Schematic view of subnuclear structures.** A typical vertebrate nucleus has several Cajal bodies, which are proposed to be the sites where snRNPs and snoRNPs undergo their final modifications. Interchromatin granule clusters are proposed to be storage sites for fully mature snRNPs. A typical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

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

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

## Summary

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

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

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

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

7th

# FROM RNA TO PROTEIN

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