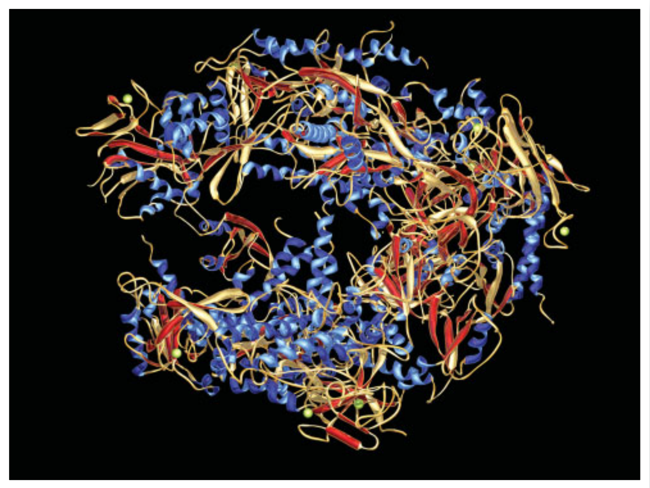
11 Transcription and RNA Processing

CHAPTER OUTLINE

- Transfer of Genetic Information: The Central Dogma
- The Process of Gene Expression
- Transcription in Prokaryotes
- Transcription and RNA Processing in Eukaryotes
- Interrupted Genes in Eukaryotes: Exons and Introns
- Removal of Intron Sequences by RNA Splicing

Storage and Transmission of Information with Simple Codes

We live in the age of the computer. It has an impact on virtually all aspects of our lives, from driving to work to watching spaceships land on the moon. These electronic wizards can store, retrieve, and analyze data with lightning-like speed. The "brain" of the computer is a small chip of silicon, the microprocessor, which contains a sophisticated and integrated array of electronic circuits capable of responding almost instantaneously to coded bursts of electrical energy. In carrying out its amazing feats, the computer uses a binary code, a language based on 0's and 1's. Thus, the alphabet used by computers is like that of the Morse code (dots and dashes) used in telegraphy. Both consist of only two symbols—in marked contrast to the 26 letters of the English alphabet. Obviously, if the computer can perform its wizardry with a binary alphabet, vast amounts of information can be stored and retrieved without using complex codes or lengthy alphabets. In this and the following chapter, we examine (1) how the genetic information of living creatures is written in an alphabet with just four letters, the four base pairs in DNA, and (2) how this genetic information is expressed during the growth and development of an organism. We will see that RNA plays a key role in the process of gene expression.



Computer model of the structure of RNA polymerase II, which catalyzes transcription of nuclear genes in eukaryotes.

Transfer of Genetic Information: The Central Dogma

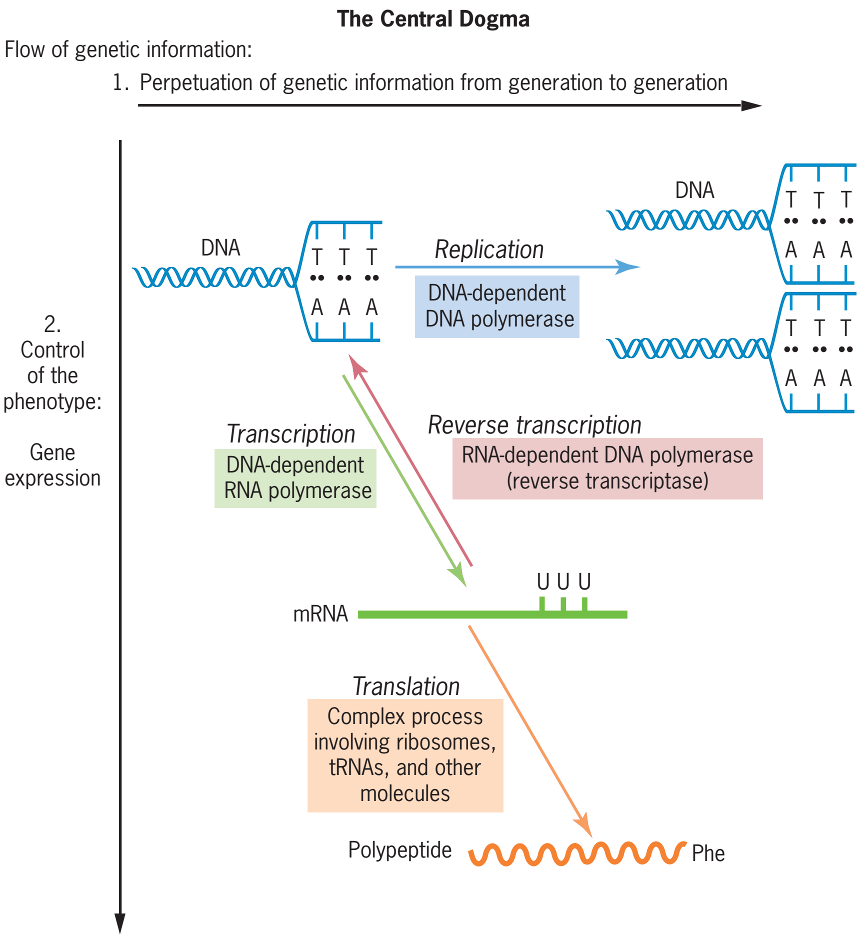
The central dogma of biology is that information stored in DNA is transferred to RNA molecules during transcription and to proteins during translation.

According to the central dogma of molecular biology, genetic information usually flows (1) from DNA to DNA during its transmission from generation to generation and (2) from DNA to protein during its phenotypic expression in an organism (■ **Figure 11.1**). During the replication of RNA viruses, information is also transmitted from RNA to RNA. The transfer of genetic information from DNA to protein involves two steps: (1) **transcription,** the transfer of the genetic information from DNA to RNA, and (2) **translation,** the transfer of information from RNA to protein. In addition, genetic information flows from RNA to DNA during the conversion of the genomes of RNA tumor viruses to their DNA proviral forms (Chapter 21). Thus, the transfer of genetic information from DNA to RNA is sometimes reversible, whereas the transfer of information from RNA to protein is always irreversible.

TRANSCRIPTION AND TRANSLATION

As we discussed earlier, the expression of genetic information occurs in two steps: transcription and translation (Figure 11.1). During transcription, one strand of DNA of a gene is used as a template to synthesize a complementary strand of RNA, called the gene **transcript.** For example, in Figure 11.1, the DNA strand containing the nucleotide sequence AAA is used as a template to produce the complementary sequence UUU in the RNA transcript. During translation, the sequence of nucleotides in the RNA transcript is converted into the sequence of amino acids in the polypeptide gene product. This conversion is governed by the **genetic code,** the specification of amino acids by nucleotide triplets called **codons** in the gene transcript. For example, the UUU triplet in the RNA transcript shown in Figure 11.1 specifies the amino acid phenylalanine (Phe) in the polypeptide gene product. Translation takes place on intricate macromolecular machines called **ribosomes,** which are composed of three to five RNA molecules and 50 to 90 different proteins. However, the process of translation also requires the participation of many other macromolecules. This chapter focuses on transcription; translation is the subject of Chapter 12.

The RNA molecules that are translated on ribosomes are called **messenger RNAs (mRNAs).** In prokaryotes, the product of transcription, the **primary transcript,** usually is equivalent to the mRNA molecule (■ **Figure 11.2*a***). In eukaryotes, primary transcripts often must be processed by the excision of specific sequences and the modification of both termini before they can be translated (■ **Figure 11.2*b***). Thus, in eukaryotes, primary transcripts usually are precursors to mRNAs and, as such, are called **pre-mRNAs.** Most of the nuclear genes in higher eukaryotes and some in lower eukaryotes contain noncoding sequences called *introns* that separate the expressed sequences or *exons* of these genes. The entire sequences of these *split genes* are transcribed into pre-mRNAs, and the noncoding intron sequences are subsequently removed by *splicing reactions* carried out on macromolecular structures called *spliceosomes*.


The Central Dogma
Flow of genetic information:
1. Perpetuation of genetic information from generation to generation
2. Control of the phenotype:
Gene expression
DNA
T T T
A A A
Replication
DNA-dependent DNA polymerase
DNA
T T T
A A A
T T T
A A A
Transcription
DNA-dependent RNA polymerase
Reverse transcription
RNA-dependent DNA polymerase (reverse transcriptase)
mRNA
U U U
Translation
Complex process involving ribosomes, tRNAs, and other molecules
Polypeptide
Phe

■ **FIGURE 11.1** The flow of genetic information according to the central dogma of molecular biology. Replication, transcription, and translation occur in all organisms; reverse transcription occurs in cells infected with certain RNA viruses. Not shown is the transfer of information from RNA to RNA during the replication of RNA viruses.

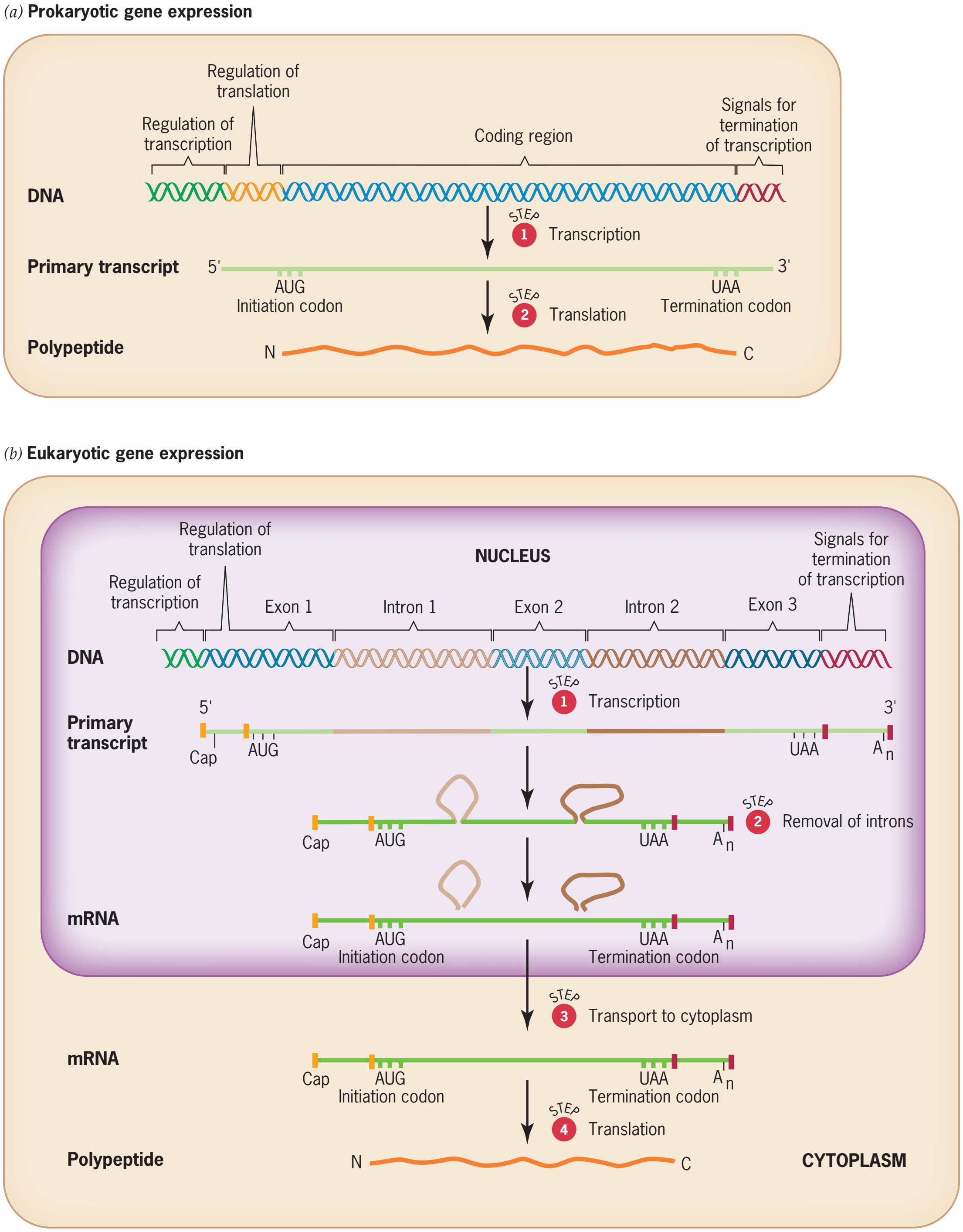


FIGURE 11.2 Gene expression involves two steps: transcription and translation, in both prokaryotes (*a*) and eukaryotes (*b*). In eukaryotes, the primary transcripts or pre-mRNAs often must be processed by the excision of introns and the addition of 5′ 7-methyl guanosine caps (CAP) and 3′ poly(A) tails $[(A)_n]$. In addition, eukaryotic mRNAs must be transported from the nucleus to the cytoplasm where they are translated.

FIVE TYPES OF RNA MOLECULES

Five different classes of RNA molecules play essential roles in gene expression. We have already discussed messenger RNAs, the intermediaries that carry genetic information from DNA to the ribosomes where proteins are synthesized. **Transfer RNAs (tRNAs)** are small RNA molecules that function as adaptors between amino acids and the codons in mRNA during translation. **Ribosomal RNAs (rRNAs)** are structural and catalytic components of the ribosomes, the intricate machines that translate nucleotide sequences of mRNAs into amino acid sequences of polypeptides. **Small nuclear RNAs (snRNAs)** are structural components of spliceosomes, the nuclear organelles that excise introns from gene transcripts. **Micro RNAs (miRNAs)** are short 20- to 22-nucleotide single-stranded RNAs that are cleaved from small hairpin-shaped precursors and block the expression of complementary or partially complementary mRNAs by either causing their degradation or repressing their translation. The roles of mRNAs and snRNAs are discussed in this chapter. The structures and functions of tRNAs and rRNAs will be discussed in detail in Chapter 12. The mechanisms by which miRNAs regulate gene expression are discussed in Chapter 19.

All five types of RNA—mRNA, tRNA, rRNA, snRNA, and miRNA—are produced by transcription. Unlike mRNAs, which specify polypeptides, the final products of tRNA, rRNA, snRNA, and miRNA genes are RNA molecules. Transfer RNA, ribosomal RNA, snRNA, and miRNA molecules are not translated. ■ **Figure 11.3** shows an overview of gene expression in eukaryotes, emphasizing the transcriptional origin and functions of the five types of RNA molecules. The process is similar in prokaryotes. However, in prokaryotes, the DNA is not separated from the ribosomes by a nuclear envelope. In addition, prokaryotic genes seldom contain noncoding sequences that are removed during RNA transcript processing.

KEY POINTS

- *The central dogma of molecular biology is that genetic information flows from DNA to DNA during chromosome replication, from DNA to RNA during transcription, and from RNA to protein during translation.*
- *Transcription involves the synthesis of an RNA transcript complementary to one strand of DNA of a gene.*
- *Translation is the conversion of information stored in the sequence of nucleotides in the RNA transcript into the sequence of amino acids in the polypeptide gene product, according to the specifications of the genetic code.*

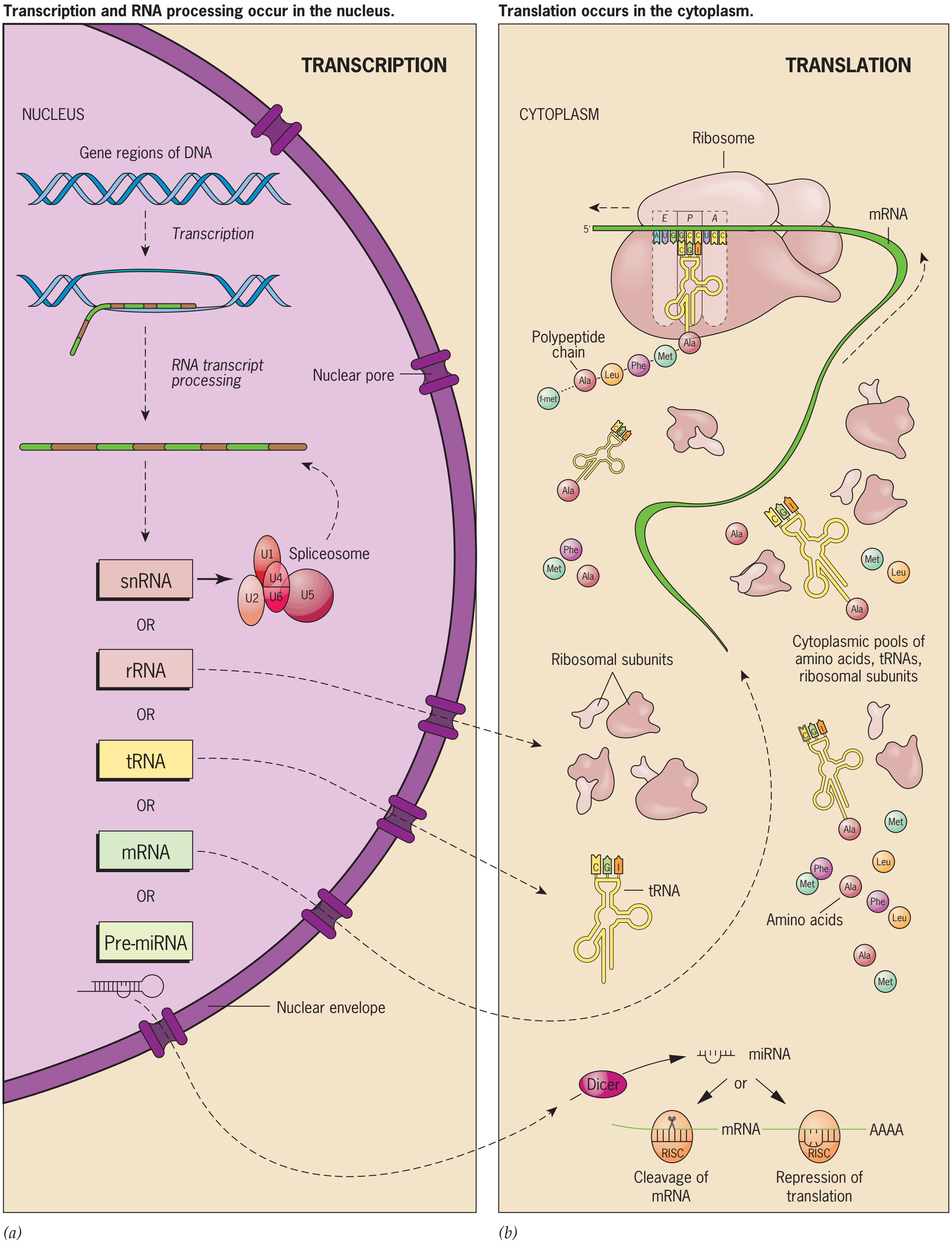
The Process of Gene Expression

Information stored in the nucleotide sequences of genes is translated into the amino acid sequences of proteins through unstable intermediaries called messenger RNAs.

How do genes control the phenotype of an organism? How do the nucleotide sequences of genes direct the growth and development of a cell, a tissue, an organ, or an entire living creature? Geneticists know that the phenotype of an organism is produced by the combined effects of all its genes acting within the constraints imposed by the environment. They also know that the number of genes in an organism varies over an enormous range, with gene number increasing with the developmental complexity of the species. The RNA genomes of the smallest viruses such as phage MS2 contain only four genes, whereas large viruses such as phage T4 have about 200 genes. Bacteria such as *E. coli* have approximately 4000 genes, and mammals, including humans, have about 20,500 genes. In this and the following chapter, we focus on the mechanisms by which genes direct the synthesis of their products, namely, RNAs and proteins. The mechanisms by which these gene products collectively control the phenotypes of mature organisms are discussed in subsequent chapters, especially Chapter 20.

AN mRNA INTERMEDIARY

If most of the genes of a eukaryote are located in the nucleus, and if proteins are synthesized in the cytoplasm, how do these genes control the amino acid sequences of their protein products? The genetic information stored in the sequences of nucleotide

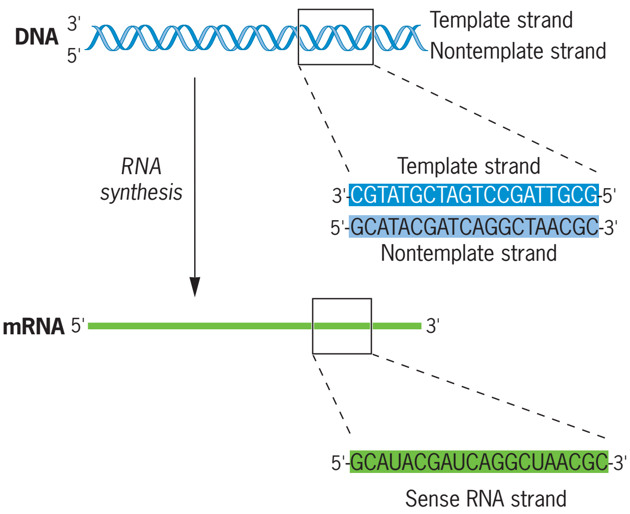


Transcription and RNA processing occur in the nucleus.

Translation occurs in the cytoplasm.

■ **FIGURE 11.3** An overview of gene expression, emphasizing the transcriptional origin of miRNA, snRNA, tRNA, rRNA, and mRNA, the splicing function of snRNA, the regulation of gene expression by miRNA, and the translational roles of tRNA, rRNA, mRNA, and ribosomes. Dicer is a nuclease that processes the miRNA precursor into miRNA, and RISC is the *R*NA-*i*nduced *s*ilencing *c*omplex.

pairs in genes must somehow be transferred to the sites of protein synthesis in the cytoplasm. Messengers are needed to transfer genetic information from the nucleus to the cytoplasm. Although the need for such messengers is most obvious in eukaryotes, the first evidence for their existence came from studies of prokaryotes. Some of the early evidence for the existence of short-lived messenger RNAs is discussed in Appendix D: Evidence for an Unstable Messenger RNA.



■ **FIGURE 11.4** RNA synthesis utilizes only one DNA strand of a gene as template.

GENERAL FEATURES OF RNA SYNTHESIS

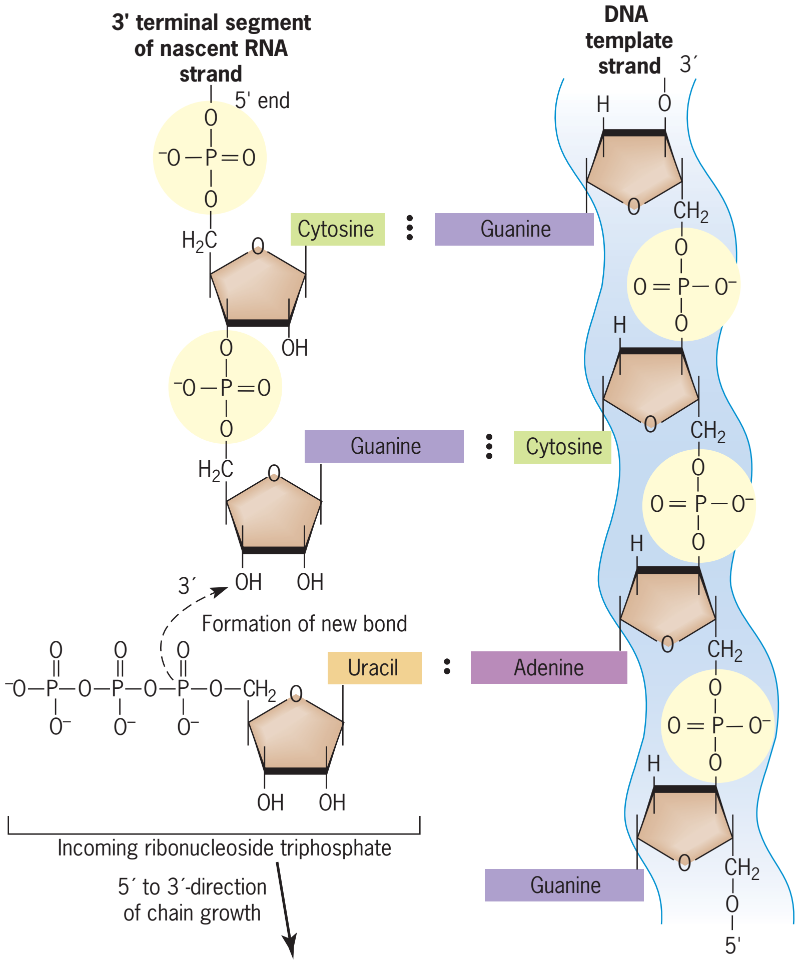
RNA synthesis occurs by a mechanism that is similar to that of DNA synthesis (Chapter 10) except that (1) the precursors are *ribonucleoside triphosphates* rather than deoxyribonucleoside triphosphates, (2) only one strand of DNA is used as a template for the synthesis of a complementary RNA chain in any given region, and (3) RNA chains can be initiated *de novo*, without any requirement for a preexisting primer strand. The RNA molecule produced will be complementary and antiparallel to the DNA **template strand** and identical, except that uridine residues replace thymidines, to the DNA **nontemplate strand** (■ **Figure 11.4**). If the RNA molecule is an mRNA, it will specify amino acids in the protein gene product. Therefore, mRNA molecules are coding strands of RNA. They are also called **sense strands** of RNA because their nucleotide sequences "make sense" in that they specify sequences of amino acids in the protein gene products. An RNA molecule that is complementary to an mRNA is referred to as **antisense RNA.** This terminology is sometimes extended to the two strands of DNA. However, usage of the terms *sense* and *antisense* to denote DNA strands has been inconsistent. Thus, we will use *template strand* and *nontemplate strand* to refer to the transcribed and nontranscribed strands, respectively, of a gene.

The synthesis of RNA chains, like DNA chains, occurs in the 5′ → 3′ direction, with the addition of ribonucleotides to the 3′-hydroxyl group at the end of the chain (■ **Figure 11.5**). The reaction involves a nucleophilic attack by the 3′-OH on the nucleotidyl (interior) phosphorus atom of the ribonucleoside triphosphate precursor with the elimination of pyrophosphate, just as in DNA synthesis. This reaction is catalyzed by enzymes called **RNA polymerases.** The overall reaction is as follows:

$$n(\text{RTP}) \xrightarrow[\text{RNA polymerase}]{\text{DNA template}} (\text{RMP})_n + n(\text{PP})$$

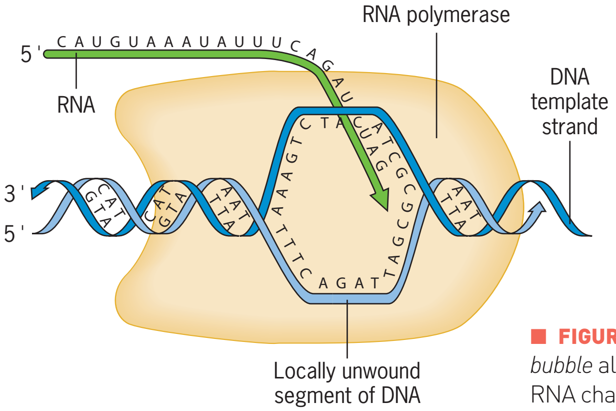
where *n* is the number of moles of ribonucleotide triphosphate (RTP) consumed, ribonucleotide monophosphate (RMP) incorporated into RNA, and pyrophosphate (PP) produced.

RNA polymerases bind to specific nucleotide sequences called **promoters,** and with the help of proteins called transcription factors, initiate the synthesis of RNA molecules at transcription start sites near the promoters. The promoters in eukaryotes are typically more complex than those of prokaryotes. A single RNA polymerase carries out all transcription in most prokaryotes, whereas five different RNA polymerases are present in eukaryotes, with each polymerase responsible for the synthesis of a distinct class of RNAs. RNA synthesis takes place within a locally unwound


3' terminal segment of nascent RNA strand
5' end
DNA template strand 3′
Cytosine
Guanine
Guanine
Cytosine
Uracil
Adenine
Guanine
3′
OH
Formation of new bond
Incoming ribonucleoside triphosphate
5′ to 3′-direction of chain growth
5'

■ **FIGURE 11.5** The RNA chain elongation reaction catalyzed by RNA polymerase.

segment of DNA, sometimes called a **transcription bubble,** which is produced by RNA polymerase (■ Figure 11.6). The nucleotide sequence of an RNA molecule is complementary to that of its DNA template strand, and RNA synthesis is governed by the same base-pairing rules as DNA synthesis, but uracil replaces thymine. As a result, the origin of RNA transcripts can be determined by studying their hybridization to DNAs from different sources such as the chromosome(s) of the cell, viruses, and other infectious organisms (see Problem-Solving Skills: Distinguishing RNAs Transcribed from Viral and Host DNAs).



■ **FIGURE 11.6** RNA synthesis occurs within a locally unwound segment of DNA. This *transcription bubble* allows a few nucleotides in the template strand to base-pair with the growing end of the RNA chain. The unwinding and rewinding of the DNA molecule are catalyzed by RNA polymerase.

PROBLEM-SOLVING SKILLS

Distinguishing RNAs Transcribed from Viral and Host DNAs

THE PROBLEM

E. coli cells that have been infected with a virus present the opportunity for the cells to make two types of RNA transcripts: bacterial and viral. If the virus is a lytic bacteriophage such as T4, only viral transcripts are made; if it is a nonlytic bacteriophage such as M13, both viral and bacterial transcripts are made; and if it is a quiescent prophage such as lambda, only bacterial transcripts are made. Suppose that you have just identified a new DNA virus. How could you determine which types of RNA transcripts are made in cells infected with this virus?

FACTS AND CONCEPTS

1. During the first step in gene expression (transcription), one strand of DNA is used as a template for the synthesis of a complementary strand of RNA.
2. RNA can be labeled with ^{3}H by growing cells in medium containing ^{3}H-uridine.
3. DNA can be denatured—separated into its constituent single strands—by exposing it to high temperature or high pH.
4. Viral DNAs and host cell DNAs can both be purified, denatured, and bound to membranes for use in subsequent hybridization experiments (see Figure 1 in Appendix D: Evidence for an Unstable Messenger RNA).
5. Under the appropriate conditions, complementary single-stranded RNA and DNA molecules will form stable double helices *in vitro*.

ANALYSIS AND SOLUTION

The source of the RNA transcripts being synthesized in virus-infected cells can be determined by incubating the infected cells for a short period of time in medium containing ^{3}H-uridine, purifying the RNA from these cells, and then hybridizing it to single-stranded viral and bacterial DNAs.

a. You should prepare one membrane with denatured viral DNA bound to it, a second membrane with denatured host DNA bound to it, and a third membrane with no DNA to serve as a control to measure nonspecific binding of ^{3}H-labeled RNA.

b. You should then prepare an appropriate hybridization solution and place the three membranes—one with viral DNA, one with host DNA, and one with no DNA—in this solution.

c. You next add a sample of the purified ^{3}H-labeled RNA and allow it to hybridize with the DNA on the membranes. Then you wash the membranes thoroughly to remove any nonhybridized RNA. The RNA that remains has either bound specifically to DNA on the membrane or it has bound nonspecifically to the membrane itself. The extent of the RNA binding can be determined by measuring how radioactive each membrane is.

d. Radioactivity on the membrane that had no DNA represents nonspecific "background" binding of RNA to the membrane. This radioactivity can be subtracted from the levels of radioactivity on the other two membranes to measure the specific binding of RNA to viral or bacterial DNA. The results will tell you whether the labeled transcripts were synthesized from viral DNA templates, bacterial DNA templates, or both. With phage T4-infected cells, phage M13-infected cells, and cells containing lambda prophages the results might be summarized as follows. (The plus signs indicate the presence of RNA transcripts that hybridize specifically.)

	RNA Hybridized to Membrane Containing	
	***E. coli* DNA**	**Phage DNA**
Phage T4-infected *E. coli* cells	–	+
Phage M13-infected *E. coli* cells	+	+
E. coli cells carrying lambda prophages	+	–

Which pattern do you observe in cells infected with the newly discovered virus?

For further discussion visit the Student Companion site.

- *In eukaryotes, genes are present in the nucleus, whereas polypeptides are synthesized in the cytoplasm.*
- *Messenger RNA molecules function as intermediaries that carry genetic information from DNA to the ribosomes, where proteins are synthesized.*
- *RNA synthesis, catalyzed by RNA polymerases, is similar to DNA synthesis in many respects.*
- *RNA synthesis occurs within a localized region of strand separation, and only one strand of DNA functions as a template for RNA synthesis.*

KEY POINTS

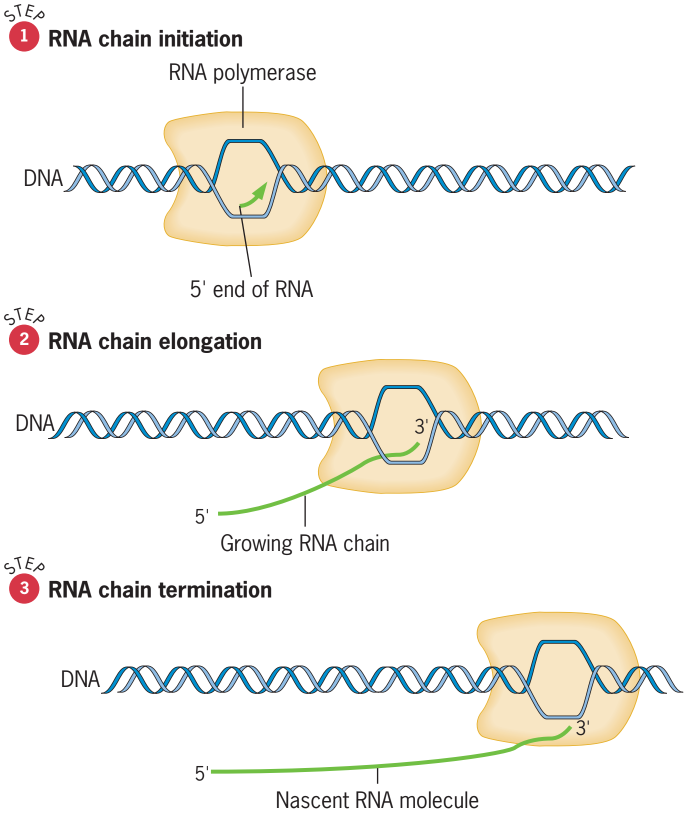
Transcription in Prokaryotes

Transcription—the first step in gene expression—transfers the genetic information stored in DNA (genes) into messenger RNA molecules that carry the information to the ribosomes—the sites of protein synthesis—in the cytoplasm.

The basic features of transcription are the same in both prokaryotes and eukaryotes, but many of the details—such as the promoter sequences—are different. The RNA polymerase of *E. coli* has been studied in great detail and will be discussed here. It catalyzes all RNA synthesis in this species. The RNA polymerases of archaea have quite different structures; they will not be discussed here.

A segment of DNA that is transcribed to produce one RNA molecule is called a **transcription unit.** Transcription units may be equivalent to individual genes, or they may include several contiguous genes. Large transcripts that carry the coding sequences of several genes are common in bacteria. The process of transcription can be divided into three stages: (1) **initiation** of a new RNA chain, (2) **elongation** of the chain, and (3) **termination** of transcription and release of the nascent RNA molecule (■ **Figure 11.7**).

When discussing transcription, biologists often use the terms *upstream* and *downstream* to refer to regions located toward the 5′ end and the 3′ end, respectively, of the transcript from some site in the mRNA molecule. These terms are based on the fact that RNA synthesis always occurs in the 5′ to 3′ direction. Upstream and downstream regions of genes are the DNA sequences specifying the corresponding 5′ and 3′ segments of their transcripts relative to a specific reference point.

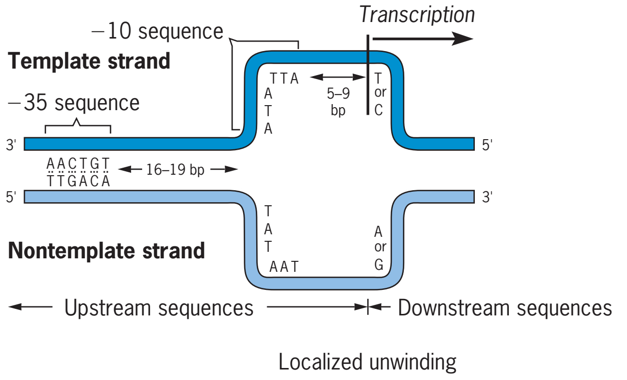

STEP 1 RNA chain initiation
RNA polymerase
DNA
5' end of RNA
STEP 2 RNA chain elongation
DNA
3'
5'
Growing RNA chain
STEP 3 RNA chain termination
DNA
3'
5'
Nascent RNA molecule

■ **FIGURE 11.7** The three stages of transcription: initiation, elongation, and termination.

RNA POLYMERASES: COMPLEX ENZYMES

The RNA polymerases that catalyze transcription are complex, multimeric proteins. The *E. coli* RNA polymerase has a molecular weight of about 480,000 and consists of five polypeptides. Two of these are identical; thus, the enzyme contains four distinct polypeptides. The complete RNA polymerase molecule, the **holoenzyme,** has the composition $\alpha_2\beta\beta'\sigma$. The α subunits are involved in the assembly of the *tetrameric core* ($\alpha_2\beta\beta'$) of RNA polymerase. The β subunit contains the ribonucleoside triphosphate binding site, and the β′ subunit harbors the DNA template-binding region.

One subunit, the **sigma (σ) factor,** is involved only in the initiation of transcription; it plays no role in chain elongation. After RNA chain initiation has occurred, the σ factor is released, and chain elongation (see Figure 11.5) is catalyzed by the core enzyme ($\alpha_2\beta\beta'$). The function of sigma is to recognize and bind RNA polymerase to the transcription initiation or promoter sites in DNA. The core enzyme (with no σ) will catalyze RNA synthesis from DNA templates *in vitro*, but, in so doing, it will initiate RNA chains at random sites on both strands of DNA. In contrast, the holoenzyme (σ present) initiates RNA chains *in vitro* only at sites used *in vivo*.



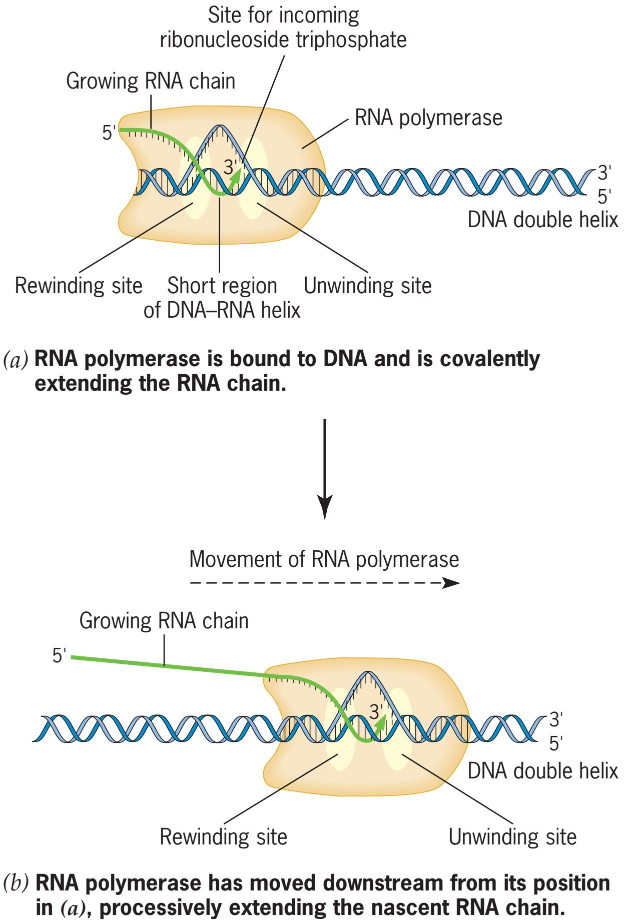
■ **FIGURE 11.8** Structure of a typical promoter in *E. coli*. RNA polymerase binds to the −35 sequence of the promoter and initiates unwinding of the DNA strands at the AT-rich −10 sequence. Transcription begins within the transcription bubble at a site five to nine base pairs beyond the −10 sequence.

INITIATION OF RNA CHAINS

Initiation of RNA chains involves three steps: (1) binding of the RNA polymerase holoenzyme to a promoter region in DNA; (2) the localized unwinding of the two strands of DNA by RNA polymerase, providing a template strand free to base-pair with incoming ribonucleotides; and (3) the formation of phosphodiester bonds between the first few ribonucleotides in the nascent RNA chain. The holoenzyme remains bound at the promoter region during the synthesis of the first eight or nine bonds; then the sigma factor is released, and the core enzyme begins the elongation phase of RNA synthesis. During initiation, short chains of two to nine ribonucleotides are synthesized and released. This abortive synthesis stops once chains of 10 or more ribonucleotides have been synthesized and RNA polymerase has begun to move downstream from the promoter.

By convention, the nucleotide pairs or nucleotides within and adjacent to transcription units are numbered relative to the transcript initiation site (designated +1)—the nucleotide pair corresponding to the first (5′) nucleotide of the RNA transcript. Base pairs preceding the initiation site are given minus (−) prefixes; those following (relative to the direction of transcription) the initiation site are given plus (+) prefixes. Nucleotide sequences preceding the initiation site are referred to as **upstream sequences;** those following the initiation site are called **downstream sequences.**

As mentioned earlier, the sigma subunit of RNA polymerase mediates its binding to promoters in DNA. Hundreds of *E. coli* promoters have been sequenced and found to have surprisingly little in common. Two short sequences within these promoters are sufficiently conserved to be recognized, but even these are seldom identical in two different promoters. The midpoints of the two conserved sequences occur at about 10 and 35 nucleotide pairs, respectively, before the transcription-initiation site (■ Figure 11.8). Thus they are called the **−10 sequence** and the **−35 sequence,** respectively. Although these sequences vary slightly from gene to gene, some nucleotides are highly conserved. The nucleotide sequences that are present in such conserved genetic elements most often are called **consensus sequences.** The −10 consensus sequence in the nontemplate strand is TATAAT; the −35 consensus sequence is TTGACA. The sigma subunit initially recognizes and binds to the −35 sequence; thus, this sequence is sometimes called the **recognition sequence.** The AT-rich −10 sequence facilitates the localized unwinding of DNA, which is an essential prerequisite to the synthesis of a new RNA chain. The distance between the −35 and −10 sequences is highly conserved in *E. coli* promoters, never being less than 15 or more than 20 nucleotide pairs in length. In addition, the first or 5′ base in *E. coli* RNAs is usually (>90 percent) a purine.



(a) **RNA polymerase is bound to DNA and is covalently extending the RNA chain.**

(b) **RNA polymerase has moved downstream from its position in *(a)*, processively extending the nascent RNA chain.**

■ **FIGURE 11.9** Elongation of an RNA chain catalyzed by RNA polymerase in *E. coli*.

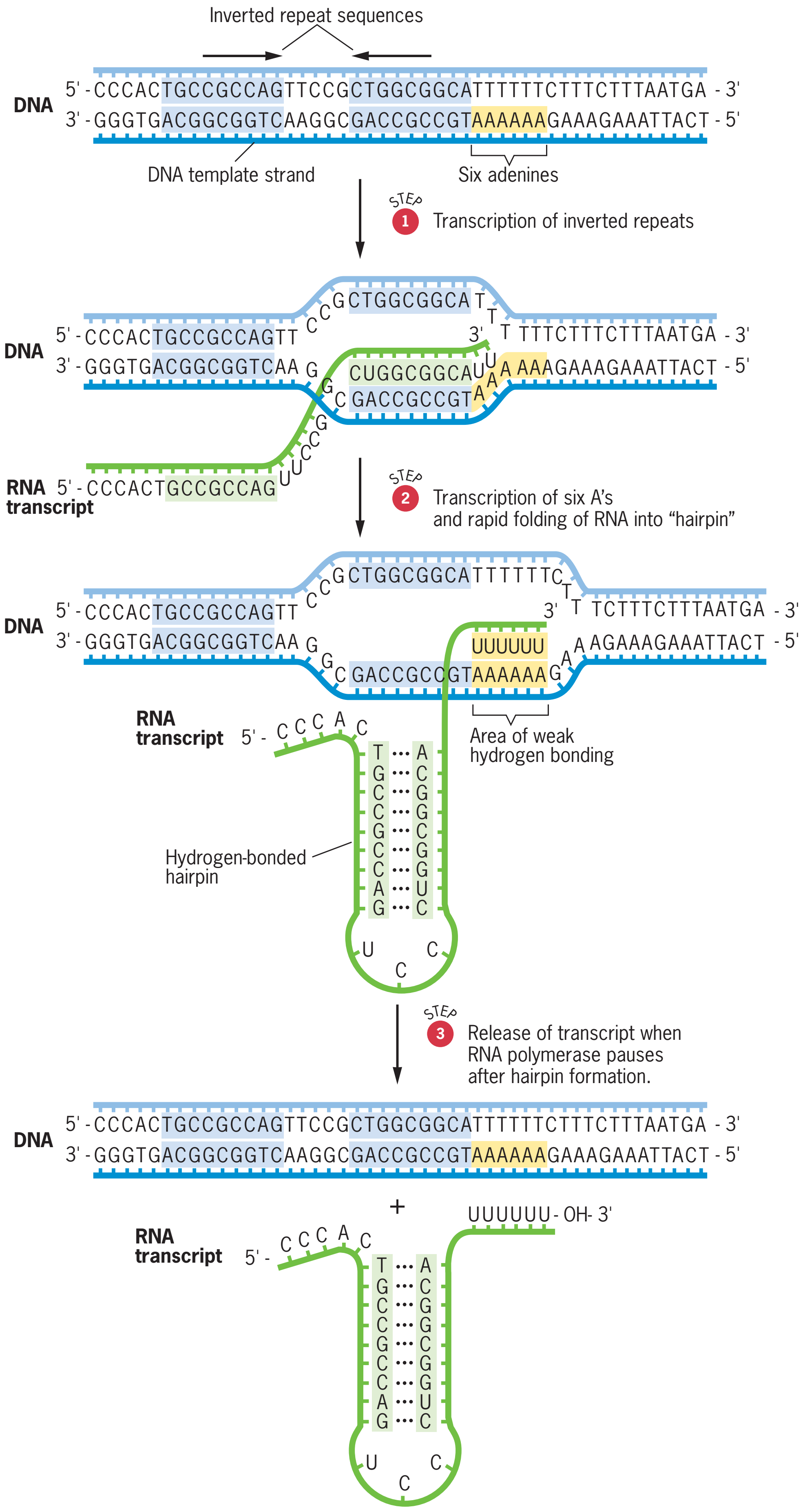
ELONGATION OF RNA CHAINS

Elongation of RNA chains is catalyzed by the RNA polymerase core enzyme, after the release of the σ subunit. The covalent extension of RNA chains (see Figure 11.5) takes place within the transcription bubble, a locally unwound segment of DNA. The RNA polymerase molecule contains both DNA unwinding and DNA rewinding activities. RNA polymerase continuously unwinds the DNA double helix ahead of the polymerization site and rewinds the complementary DNA strands behind the polymerization site as it moves along the double helix (■ Figure 11.9). In *E. coli*, the average length of a transcription bubble is 18 nucleotide pairs, and about 40 ribonucleotides are incorporated into the growing RNA chain per second. The nascent RNA chain is displaced from the DNA template strand as RNA polymerase moves along the DNA molecule. The region of transient base-pairing between the growing chain and the DNA template strand is very short, perhaps only three base pairs in length. The stability of the transcription complex is maintained primarily by the binding of the DNA and the growing RNA chain to RNA polymerase, rather than by the base-pairing between the template strand of DNA and the nascent RNA.

TERMINATION OF RNA CHAINS

Termination of RNA chains occurs when RNA polymerase encounters a **termination signal.** When it does, the transcription complex dissociates, releasing the nascent RNA molecule. There are two types of transcription terminators in *E. coli*. One type results in termination only in the presence of a protein called *rho* (ρ); therefore, such termination sequences are called *rho-dependent terminators*. The other type results in the termination of transcription without the involvement of rho; such sequences are called *rho-independent terminators*.

Rho-independent terminators contain a GC-rich region followed by six or more AT base pairs, with the A's present in the template strand (■ **Figure 11.10**, top). The nucleotide sequence of the GC-rich region contains inverted repeats—sequences of nucleotides in each DNA strand that are inverted and complementary. When transcribed, these inverted repeat regions produce single-stranded RNA sequences that can base-pair and form hairpin structures (Figure 11.10, bottom). The RNA hairpin structures form immediately after the synthesis of the participating regions of the RNA chain and retard the movement of RNA polymerase molecules along the DNA, causing pauses in chain extension. Since AU base-pairing is weak, requiring less energy


Inverted repeat sequences
DNA
5' - CCCACTGCCGCCAGTTCCGCTGGCGGCATTTTTTCTTTCTTTAATGA - 3'
3' - GGGTGACGGCGGTCAAGGCGACCGCCGTAAAAAAGAAAGAAATTACT - 5'
DNA template strand
Six adenines
STEP 1 Transcription of inverted repeats
DNA
5' - CCCACTGCCGCCAGTT
CC GCTGGCGGCA T T
3' T TTTCTTTCTTTAATGA - 3'
3' - GGGTGACGGCGGTCAA G
G C GACCGCCGT A
CUGGCGGCAU U A A AAAGAAAGAAATTACT - 5'
RNA transcript 5' - CCCACTGCCGCCAG UU CC G G
STEP 2 Transcription of six A's and rapid folding of RNA into "hairpin"
DNA
5' - CCCACTGCCGCCAGTT
CC GCTGGCGGCATTTTTTC T
3' T TCTTTCTTTAATGA - 3'
3' - GGGTGACGGCGGTCAA G
G C GACCGCCGTAAAAAA G A
UUUUUU
A AGAAAGAAATTACT - 5'
RNA transcript 5' - C C C A C
T G C C G C C A G
A C G G C G G U C
U C C
Area of weak hydrogen bonding
Hydrogen-bonded hairpin
STEP 3 Release of transcript when RNA polymerase pauses after hairpin formation.
DNA
5' - CCCACTGCCGCCAGTTCCGCTGGCGGCATTTTTTCTTTCTTTAATGA - 3'
3' - GGGTGACGGCGGTCAAGGCGACCGCCGTAAAAAAGAAAGAAATTACT - 5'
+
UUUUUU - OH- 3'
RNA transcript 5' - C C C A C
T G C C G C C A G
A C G G C G G U C
U C C

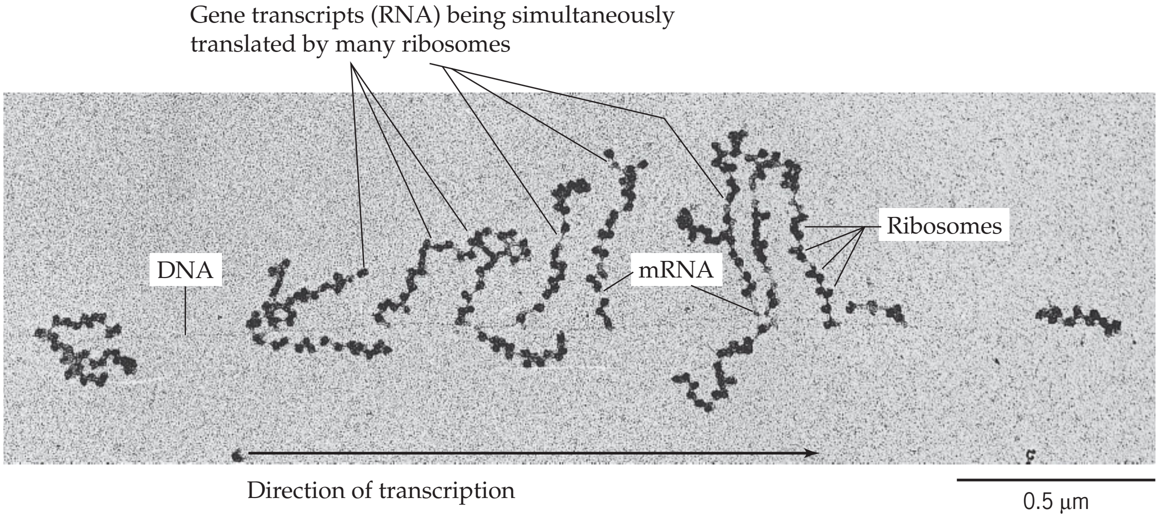
■ **FIGURE 11.10** Mechanism of rho-independent termination of transcription. As transcription proceeds along a DNA template, a region of DNA is encountered that contains inverted repeat sequences (shaded). When these repeat sequences are transcribed, the RNA transcript will contain sequences that are complementary to each other. As a result, they will hydrogen bond and form a hairpin structure. When RNA polymerase encounters this hairpin, it will pause, and the weak hydrogen bonds between the A's that follow in the template strand and the U's in the newly synthesized transcript will break, releasing the transcript from the DNA.

to separate the bases than any of the other standard base pairs, the run of U's after the hairpin region facilitates the release of the newly synthesized RNA chains from the DNA template when the hairpin structure causes RNA polymerase to pause at this site.

The mechanism by which rho-dependent termination of transcription occurs is similar to that of rho-independent termination in that both involve the formation of a hydrogen-bonded hairpin structure upstream from the site of termination. In both cases, these hairpins impede the movement of RNA polymerase, causing it to pause. However, rho-dependent terminators contain two additional sequences: a 50–90 nucleotide-pair sequence upstream from the inverted repeat sequences that produces an RNA strand with many C's but few G's, which therefore forms no hairpins or other secondary structures, and a sequence specifying a rho protein binding site called *rut* (for rho utilization) near the 3′ end of the transcript. Rho protein binds to the *rut* sequence in the transcript and moves 5′ to 3′ following RNA polymerase. When polymerase encounters the hairpin, it pauses, allowing rho to catch up, pass through the hairpin, and use its helicase activity to unwind the DNA/RNA base-pairing at the terminus and release the RNA transcript.

CONCURRENT TRANSCRIPTION, TRANSLATION, AND mRNA DEGRADATION

In prokaryotes, the translation and degradation of an mRNA molecule often begin before its synthesis (transcription) is complete. Since mRNA molecules are synthesized, translated, and degraded in the 5′ to 3′ direction, all three processes can occur simultaneously on the same RNA molecule. In prokaryotes, the polypeptide-synthesizing machinery is not separated by a nuclear envelope from the site of mRNA synthesis. Therefore, once the 5′ end of an mRNA has been synthesized, it can immediately be used as a template for polypeptide synthesis. Indeed, transcription and translation often are tightly coupled in prokaryotes. Oscar Miller, Barbara Hamkalo, and colleagues developed techniques that allowed them to visualize this coupling between transcription and translation in bacteria by electron microscopy. One of their photographs showing the coupled transcription of a gene and translation of its mRNA product in *E. coli* is reproduced in ■ **Figure 11.11**.


Gene transcripts (RNA) being simultaneously translated by many ribosomes
Ribosomes
DNA
mRNA
Direction of transcription
0.5 μm

■ **FIGURE 11.11** Electron micrograph prepared by Oscar Miller and Barbara Hamkalo showing the coupled transcription and translation of a gene in *E. coli*. DNA, mRNAs, and the ribosomes translating individual mRNA molecules are visible. The nascent polypeptide chains being synthesized on the ribosomes are not visible as they fold into their three-dimensional configuration during synthesis.

KEY POINTS

- *RNA synthesis occurs in three stages: (1) initiation, (2) elongation, and (3) termination.*
- *RNA polymerases—the enzymes that catalyze transcription—are complex multimeric proteins.*
- *The covalent extension of RNA chains occurs within locally unwound segments of DNA.*
- *Chain elongation stops when RNA polymerase encounters a transcription–termination signal.*
- *Transcription, translation, and degradation of mRNA molecules often occur simultaneously in prokaryotes.*

Transcription and RNA Processing in Eukaryotes

Five different enzymes catalyze transcription in eukaryotes, and the resulting RNA transcripts undergo three important modifications, including the excision of noncoding sequences called introns. The nucleotide sequences of some RNA transcripts are modified posttranscriptionally by RNA editing.

Although the overall process of RNA synthesis is similar in prokaryotes and eukaryotes, the process is considerably more complex in eukaryotes. In eukaryotes, RNA is synthesized in the nucleus, and most RNAs that encode proteins must be transported to the cytoplasm for translation on ribosomes. There is evidence suggesting that some translation occurs in the nucleus; however, the vast majority clearly occurs in the cytoplasm.

Prokaryotic mRNAs often contain the coding regions of two or more genes; such mRNAs are said to be multigenic. In contrast, many of the eukaryotic transcripts that have been characterized contain the coding region of a single gene (are monogenic). Nevertheless, up to one-fourth of the transcription units in the small worm *Caenorhabditis elegans* may be multigenic. Clearly, eukaryotic mRNAs may be either monogenic or multigenic.

Five different RNA polymerases are present in eukaryotes, and each enzyme catalyzes the transcription of a specific class of genes. Moreover, in eukaryotes, the majority of the primary transcripts of genes that encode polypeptides undergo three major modifications prior to their transport to the cytoplasm for translation (■ **Figure 11.12**).

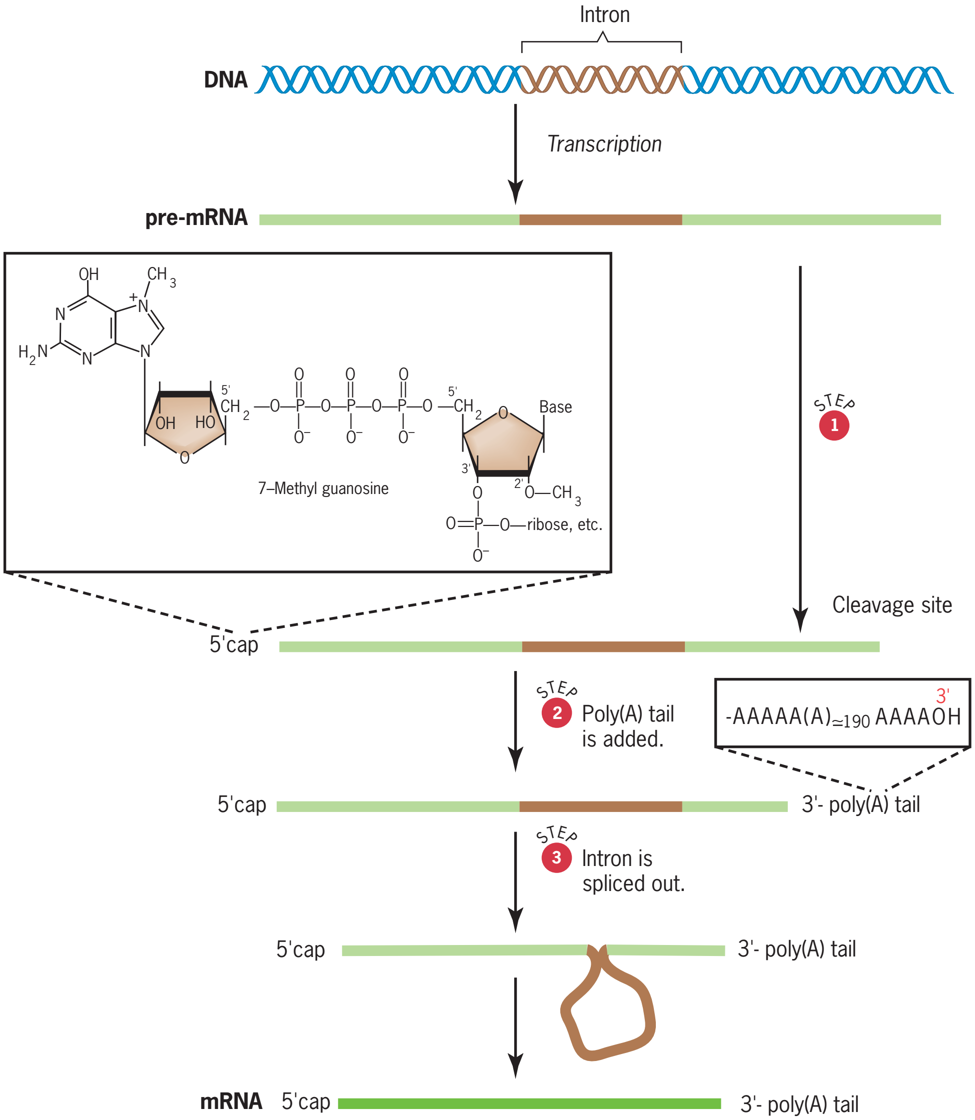
1. 7-Methyl guanosine caps are added to the 5′ ends of the primary transcripts.
2. Poly(A) tails are added to the 3′ ends of the transcripts, which are generated by cleavage rather than by termination of chain extension.
3. When present, intron sequences are spliced out of transcripts.

The **5′ cap** on most eukaryotic mRNAs is a 7-methyl guanosine residue joined to the initial nucleoside of the transcript by a 5′-5′ phosphate linkage. The 3′ **poly(A) tail** is a polyadenosine tract 20 to 200 nucleotides long.

In eukaryotes, the population of primary transcripts in a nucleus is called **heterogeneous nuclear RNA (hnRNA)** because of the large variation in the sizes of the RNA molecules present. Major portions of these hnRNAs are noncoding intron sequences, which are excised from the primary transcripts and degraded in the nucleus. Thus, much of the hnRNA actually consists of pre-mRNA molecules undergoing various processing events before leaving the nucleus. Also, in eukaryotes, RNA transcripts are coated with RNA-binding proteins during or immediately after their synthesis. These proteins protect gene transcripts from degradation by ribonucleases, enzymes that degrade RNA molecules, during processing and transport to the cytoplasm. The average half-life of a gene transcript in eukaryotes is about five hours, in contrast to an average half-life of less than five minutes in *E. coli*. This enhanced stability of gene transcripts in eukaryotes is provided, at least in part, by their interactions with RNA-binding proteins.

FIVE RNA POLYMERASES/FIVE SETS OF GENES

Whereas a single RNA polymerase catalyzes all transcription in *E. coli*, eukaryotes ranging in complexity from the single-celled yeasts to humans contain from three to five different RNA polymerases. Three enzymes, designated **RNA polymerases I, II,** and **III,** are known to be present in most, if not all, eukaryotes. All three are more



■ **FIGURE 11.12** In eukaryotes, most gene transcripts undergo three different types of posttranscriptional processing.

complex, with 10 or more subunits, than the *E. coli* RNA polymerase. Moreover, unlike the *E. coli* enzyme, all eukaryotic RNA polymerases require the assistance of other proteins called **transcription factors** in order to initiate the synthesis of RNA chains.

The key features of the five eukaryotic RNA polymerases are summarized in **Table 11.1**. RNA polymerase I is located in the nucleolus, a distinct region of the nucleus where rRNAs are synthesized and combined with ribosomal proteins. RNA polymerase I catalyzes the synthesis of all ribosomal RNAs except the small 5S rRNA. RNA polymerase II transcribes nuclear genes that encode proteins and perhaps other genes specifying hnRNAs. RNA polymerase III catalyzes the synthesis of the transfer RNA molecules, the 5S rRNA molecules, and small nuclear RNAs. To date, **RNA polymerases IV** and **V** have been identified only in plants; however, there are hints that they may exist in other eukaryotes, especially fungi.

RNA polymerases IV and V play important roles in turning off the transcription of genes by modifying the structure of chromosomes, a process called *chromatin remodeling* (see On the Cutting Edge: Chromatin Remodeling and Gene Expression and Chapter 19). Chromatin remodeling occurs when the histone tails in nucleosomes (see Figure 9.18) are chemically modified and proteins interact with these modified groups, causing the chromatin to become more or less condensed. RNA polymerase IV synthesizes transcripts that are processed into short RNAs called *small interfering RNAs* (siRNAs) that are important regulators of gene expression (see Chapter 19). One mechanism of action involves interacting with other proteins to modify (condense or relax) chromatin structure. RNA polymerase V synthesizes a subset of siRNAs and noncoding (antisense) transcripts of genes that are regulated by siRNAs. Although the details of the process are still being worked out, it seems likely that the siRNAs interact with these noncoding transcripts and nucleosome-associated proteins—some well characterized, others unknown—to condense chromatin into structures that cannot be transcribed.

TABLE 11.1

Characteristics of the Five RNA Polymerases of Eukaryotes

Enzyme	Location	Products
RNA polymerase I	Nucleolus	Ribosomal RNAs, excluding 5S rRNA
RNA polymerase II	Nucleus	Nuclear pre-mRNAs
RNA polymerase III	Nucleus	tRNAs, 5S rRNA, and other small nuclear RNAs
RNA polymerase IV	Nucleus (plant)	Small interfering RNAs (siRNAs)
RNA polymerase V	Nucleus (plant)	Some siRNAs plus noncoding (antisense) transcripts of siRNA target genes.

ON THE CUTTING EDGE

CHROMATIN REMODELING AND GENE EXPRESSION

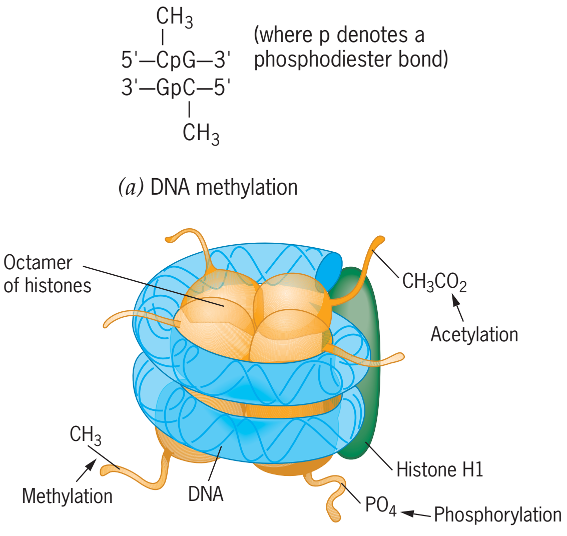
The DNA of eukaryotes is packaged into roughly 11-nm spheres called nucleosomes, which consist of DNA wound on the surface of histone octamers (see Figure 9.18). Within these nucleosomes, the charged amino-terminal tails of the histones bind tightly to DNA, keeping the structures quite compact. How, then, can the transcription factors and the large RNA polymerase complexes gain access to the promoters and transcribe the genes in nucleosomes? The answer is that the structures of nucleosomes containing genes that need to be expressed must be modified to make the promoters available to the proteins required for transcription; that is, **chromatin remodeling** must occur before transcription can begin.

There are several types of chromatin remodeling, some of which are discussed in more detail in Chapter 19. All involve chromatin-remodeling proteins, usually multimeric protein complexes. Some require the input of energy from ATP. Chromatin remodeling can occur (1) by sliding nucleosomes along DNA so that specific DNA sequences are located between nucleosomes, (2) by changing the spacing between nucleosomes, or (3) by displacing histone octamers to create nucleosome-free gaps. But what controls these chromatin-remodeling processes? What signals are required to initiate a specific pathway of chromatin remodeling?

The signals that control chromatin remodeling are still being worked out. However, chemical modifications of nucleotides in DNA and of amino acids in the protruding tails of histones in nucleosomes play key roles (■ **Figure 1**). Many of the genes of mammals contain sequences rich in the dinucleotide sequence

5′-CpG-3′
3′-GpC-5′

upstream from their transcription start sites. These CpG-rich regions are called CpG islands and are important regulatory sequences. The cytosines in the CpG islands are subject to methylation, the addition of methyl (CH$_3$) groups, and methylated CpG islands, in turn, are binding sites for proteins that regulate transcription. In many cases, the methylation of CpG islands results in the repression of transcription of nearby genes. However, in some cases, recent studies indicate that chromatin remodeling can lead to global changes in gene expression, including both repression and activation.



(*a*) DNA methylation

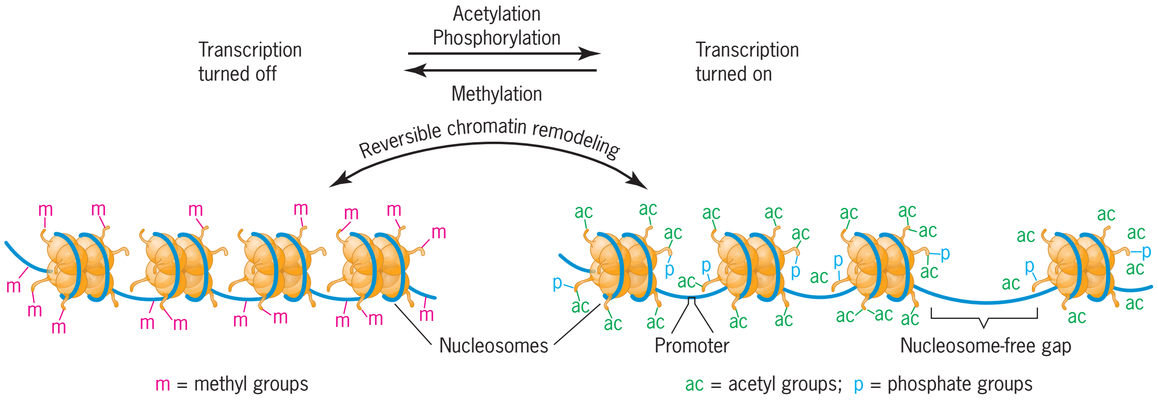
(*b*) Histone modifications

■ **FIGURE 1** Chemical modifications of (*a*) DNA and (*b*) histones involved in chromatin remodeling.

Amino acids in the protruding histone tails of nucleosomes also undergo methylation, and these methyl groups on DNA and histones work together to compact chromatin and repress gene expression. However, methylation is not the whole story! The protruding histone tails undergo two additional modifications: acetylation, the addition of acetyl (CH$_3$CO$_2$) groups; and phosphorylation, the addition of phosphate (PO$_4$) groups.

Acetylation is the more important of these modifications. Acetyl groups are added to specific lysine residues in the histone tails by enzymes called acetylases, neutralizing the positive charges of these lysines (see Figure 12.1). As a result, acetylation decreases the interaction between the negatively charged DNA and the histone tails and sets the stage for the chromatin remodeling required for the initiation of transcription.

In mammals, a protein complex called the *enhanceosome* initiates the activation process by binding to DNA upstream from the promoter and recruiting acetylase, which then adds acetyl groups to histone tails protruding from nucleosomes. Chromatin-remodeling proteins then modify the structure of the complex and make the promoter accessible to transcription factors and RNA polymerase. ■ **Figure 2** shows a



■ **FIGURE 2** A schematic overview of the effects of (1) methylation of DNA and histones, (2) acetylation of histones, and (3) phosphorylation of histones on chromatin remodeling and transcription.

continued

ON THE CUTTING EDGE *(continued)*

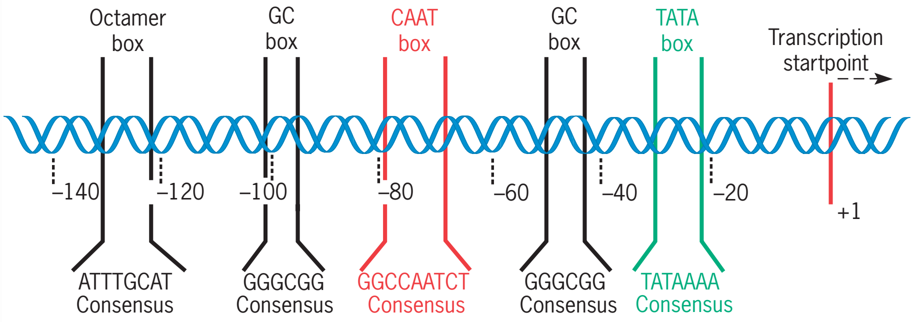
schematic overview of the effects of methylation, acetylation, and phosphorylation on chromatin remodeling and transcription.

Recent evidence indicates that global changes in gene expression—with some genes upregulated and other genes downregulated—can be caused by chromatin remodeling. Indeed, several human disorders are now known to result from genetic defects in chromatin remodeling. One form of acute lymphoblastic leukemia and Rett syndrome, a severe neurological disorder, both result from defects in chromatin remodeling. Rett syndrome, which results in loss of motor skills and mental retardation within four years of birth, is caused by mutations in the *MECP2* gene, encoding *m*ethyl-CpG-binding *p*rotein *2*, on the X chromosome. Recent evidence suggests that MECP2 is made in large quantities and binds directly to nucleosomes, actually competing with histone H1 for common binding sites. Indeed, when MECP2 binds to nucleosomes, it changes their architecture and alters the expression of genes packaged therein. Exactly how do mutations in the *MECP2* gene alter chromatin remodeling and change gene expression in neurons? The details still must be worked out. However, given the ongoing research in this field, new information will undoubtedly become available in the near future.

INITIATION OF RNA CHAINS

Unlike their prokaryotic counterparts, eukaryotic RNA polymerases cannot initiate transcription by themselves. All five eukaryotic RNA polymerases require the assistance of protein transcription factors to start the synthesis of an RNA chain. Indeed, these transcription factors must bind to a promoter region in DNA and form an appropriate initiation complex before RNA polymerase will bind and initiate transcription. Different promoters and transcription factors are utilized by RNA polymerases. In this section, we focus on the initiation of pre-mRNA synthesis by RNA polymerase II, which transcribes the vast majority of eukaryotic genes.

In all cases, the initiation of transcription involves the formation of a locally unwound segment of DNA, providing a DNA strand that is free to function as a template for the synthesis of a complementary strand of RNA (see Figure 11.6). The formation of the locally unwound segment of DNA required to initiate transcription involves the interaction of several transcription factors with specific sequences in the promoter for the transcription unit. The promoters recognized by RNA polymerase II consist of short conserved elements, or modules, located upstream from the transcription startpoint. The components of the promoter of the mouse thymidine kinase gene are shown in ■ **Figure 11.13**. Other promoters that are recognized by RNA polymerase II contain some, but not all, of these components. The conserved element closest to the transcription start site (position +1) is called the **TATA box;** it has the consensus sequence TATAAAA (reading 5′ to 3′ on the nontemplate strand) and is centered at about position −30. The TATA box plays an important role in positioning the transcription startpoint. The second conserved element is called the **CAAT box;** it usually occurs near position −80 and has the consensus sequence GGCCAATCT. Two other



■ **FIGURE 11.13** Structure of a promoter recognized by RNA polymerase II. The TATA and CAAT boxes are located at about the same positions in the promoters of most nuclear genes encoding proteins. The GC and octamer boxes may be present or absent; when present, they occur at many different locations, either singly or in multiple copies. The sequences shown here are the consensus sequences for each of the promoter elements. The conserved promoter elements are shown at their locations in the mouse thymidine kinase gene.

Solve It!

Initiation of Transcription by RNA Polymerase II in Eukaryotes

The nucleotide sequence of the nontemplate strand of a portion of the human *HBB* (β-globin) gene and the amino-terminus of its product, human β-globin (using the single-letter amino acid code), are given as follows. Remember that the nontemplate strand will have the same sequence as the transcript of the gene, but with T's in place of U's.

```
5'–CCTGTGGAGC  CACACCCTAG   GGTTGGCCAA  TCTACTCCCA
   GGAGCAGGGA  GGGCAGGAGC  CAGGGCTGGG  CATAAAAGTC
   AGGGCAGAGC  CATCTATTGC    TTACATTTGC   TTCTGACACA
   ACTGTGTTCA   CTAGCAACCT   CAAACAGACA  CCATGGTGCA
                        β-globin amino terminus:     M   V   H
  TCTGACTCCT   GAGGAGAAGT   CTGCCGTTAC  TGCCCTGTGG–3'
    L    T   P     E    E   K     S   A    V  T     A   L   W—
```

Note: Every other codon is underlined in the coding region of the gene.

Does the TATA box in this gene have the consensus sequence? If not, what is its sequence? Does this gene contain a CAAT box? Does it have the consensus sequence? Given that transcription of eukaryotic genes by RNA polymerase II almost always starts (+1 site) at an A preceded by two pyrimidines, predict the sequence of the 5′-terminus of the primary transcript of this gene.

▸ ***To see the solution to this problem, visit the Student Companion site.***

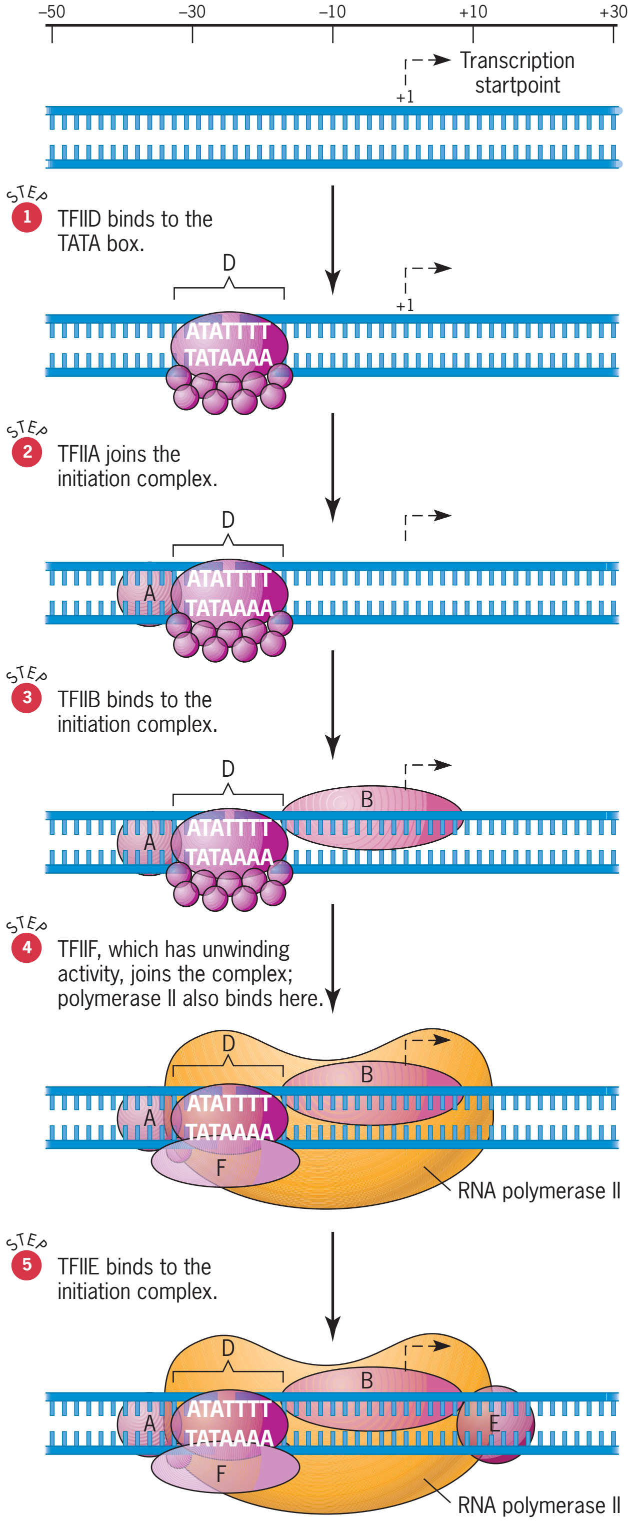
conserved elements, the *GC box*, consensus GGGCGG, and the *octamer box*, consensus ATTTGCAT, often are present in RNA polymerase II promoters; they influence the efficiency of a promoter in initiating transcription. Try Solve It: Initiation of Transcription by RNA Polymerase II in Eukaryotes to see how these conserved promoter sequences work in the human *HBB* (β-globin) gene.

The initiation of transcription by RNA polymerase II requires the assistance of several **basal transcription factors.** Still other transcription factors and regulatory sequences called *enhancers* and *silencers* modulate the efficiency of initiation (Chapter 19). The basal transcription factors must interact with promoters in the correct sequence to initiate transcription effectively (■ **Figure 11.14**). Each basal transcription factor is denoted **TFIIX** (**T**ranscription **F**actor **X** for RNA polymerase **II,** where **X** is a letter identifying the individual factor).

TFIID is the first basal transcription factor to interact with the promoter; it contains a TATA-binding protein (TBP) and several small TBP-associated proteins (Figure 11.14). Next, TFIIA joins the complex, followed by TFIIB. TFIIF first associates with RNA polymerase II, and then TFIIF and RNA polymerase II join the transcription initiation complex together. TFIIF contains two subunits, one of which has DNA-unwinding activity. Thus, TFIIF probably catalyzes the localized unwinding of the DNA double helix required to initiate transcription. TFIIE then joins the initiation complex, binding to the DNA downstream from the transcription startpoint. Two other factors, TFIIH and TFIIJ, join the complex after TFIIE, but their locations in the complex are unknown. TFIIH has helicase activity and travels with RNA polymerase II during elongation, unwinding the strands in the region of transcription (the "transcription bubble").

RNA polymerases I and III initiate transcription by processes that are similar, but somewhat simpler, than the one used by polymerase II, whereas the processes used by RNA polymerases IV and V are currently under investigation. The promoters of genes transcribed by polymerases I and III are quite different from those utilized by polymerase II, even though they sometimes contain some of the same regulatory elements. RNA polymerase I promoters are bipartite, with a core sequence extending from about −45 to +20, and an upstream control element extending from −180 to about −105. The two regions have similar sequences, and both are GC-rich. The core sequence is sufficient for initiation; however, the efficiency of initiation is strongly enhanced by the presence of the upstream control element.

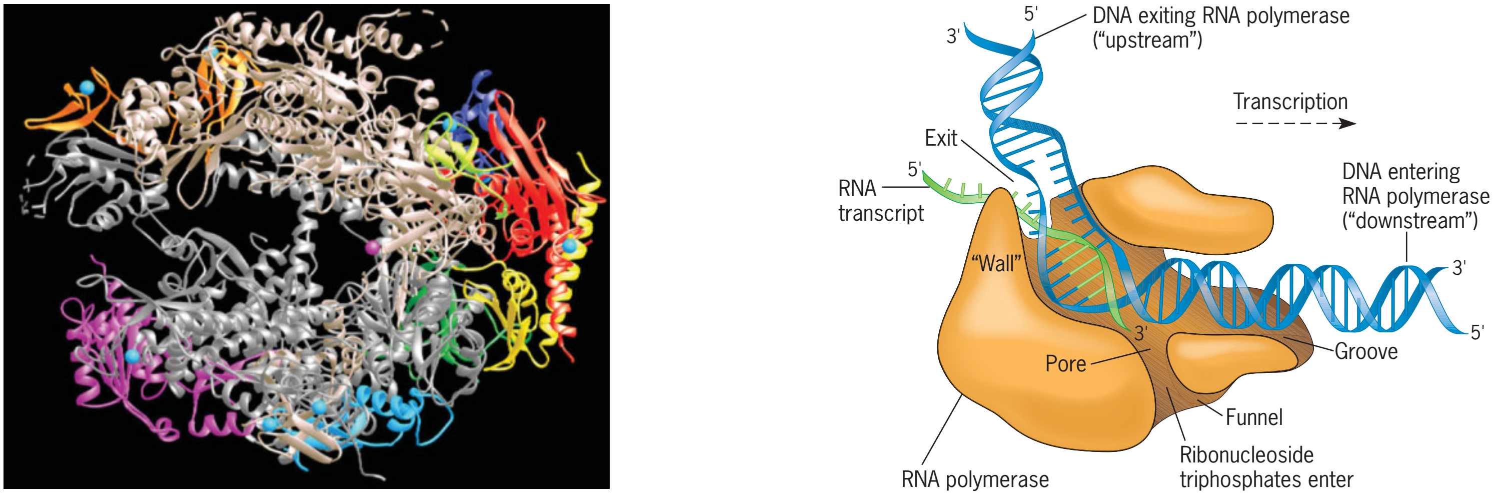
Interestingly, the promoters of most of the genes transcribed by RNA polymerase III are located within the transcription units, downstream from the transcription startpoints, rather than upstream as in units transcribed by RNA polymerases I and II. The promoters of other genes transcribed by polymerase III are located upstream of the transcription start site, just as for polymerases I and II. Actually, polymerase III promoters can be divided into three classes, two of which have promoters located within the transcription unit.


–50
–30
–10
+10
+30
Transcription startpoint
+1
STEP 1 TFIID binds to the TATA box.
D
+1
ATATTTT
TATAAAA
STEP 2 TFIIA joins the initiation complex.
D
A
ATATTTT
TATAAAA
STEP 3 TFIIB binds to the initiation complex.
D
B
A
ATATTTT
TATAAAA
STEP 4 TFIIF, which has unwinding activity, joins the complex; polymerase II also binds here.
D
B
A
ATATTTT
TATAAAA
F
RNA polymerase II
STEP 5 TFIIE binds to the initiation complex.
D
B
A
ATATTTT
TATAAAA
E
F
RNA polymerase II

■ **FIGURE 11.14** The initiation of transcription by RNA polymerase II requires the formation of a basal transcription initiation complex at the promoter region. The assembly of this complex begins when TFIID, which contains the TATA-binding protein (TBP), binds to the TATA box. The other transcription factors and RNA polymerase II join the complex in the sequence shown.

RNA CHAIN ELONGATION AND THE ADDITION OF 5′ METHYL GUANOSINE CAPS

Once eukaryotic RNA polymerases have been released from their initiation complexes, they catalyze RNA chain elongation by the same mechanism as the RNA polymerases of prokaryotes (see Figures 11.5 and 11.6). Studies on the crystal structures of various RNA polymerases have provided a good picture of key features of this important enzyme. Although the RNA polymerases of bacteria, archaea, and eukaryotes have different substructures, their


5'
3'
DNA exiting RNA polymerase ("upstream")
Transcription
Exit
5'
RNA transcript
DNA entering RNA polymerase ("downstream")
3'
"Wall"
3'
5'
Groove
Pore
Funnel
RNA polymerase
Ribonucleoside triphosphates enter

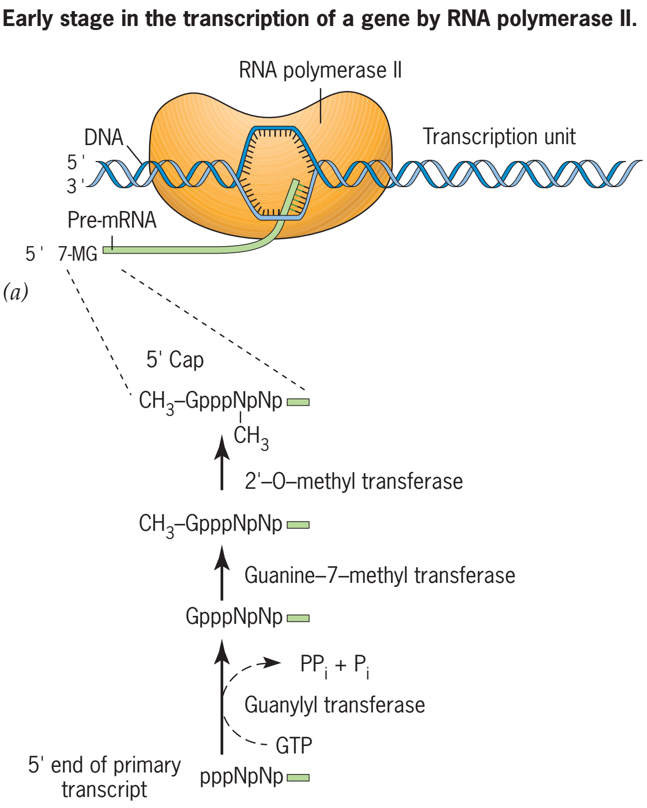
(a) **Crystal structure of yeast RNA polymerase II.**

(b) **Diagram of the interaction between DNA and RNA polymerase based on crystal structures and other structural analyses.**

■ **FIGURE 11.15** Structure of RNA polymerase. (*a*) Crystal structure of the RNA polymerase II from the yeast *S. cerevisiae*. (*b*) Diagram of an RNA polymerase, showing its interaction with DNA (blue) and the nascent RNA chain (green). Although the subunit composition of RNA polymerases varies between bacterial, archaeal, and eukaryotic enzymes, the basic structural features are quite similar in all species.

key features and mechanisms of action are quite similar. The crystal structure of RNA polymerase II (resolution = .28 nm) of *S. cerevisiae* is shown in ■ **Figure 11.15*a***. A schematic diagram showing structural features of an RNA polymerase and its interaction with DNA and the growing RNA transcript is shown in ■ **Figure 11.15*b***.

Early in the elongation process, the 5′ ends of eukaryotic pre-mRNAs are modified by the addition of 7-methyl guanosine (7-MG) caps. These 7-MG caps are added when the growing RNA chains are only about 30 nucleotides long (■ **Figure 11.16**). The 7-MG cap contains an unusual 5′-5′ triphosphate linkage (see Figure 11.12) and two or more methyl groups. These 5′ caps are added co-transcriptionally by the biosynthetic pathway shown in Figure 11.16. The 7-MG caps are recognized by protein factors involved in the initiation of translation (Chapter 12) and also help protect the growing RNA chains from degradation by nucleases.


Early stage in the transcription of a gene by RNA polymerase II.

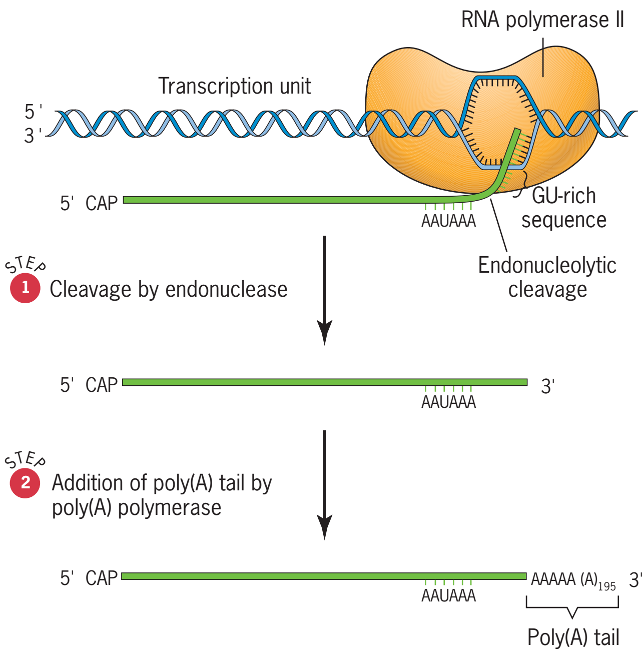
(b) **Pathway of biosynthesis of the 7-MG cap.**

■ **FIGURE 11.16** 7-Methyl guanosine (7-MG) caps are added to the 5′ ends of pre-mRNAs shortly after the elongation process begins.

Recall that eukaryotic genes are present in chromatin organized into nucleosomes (Chapter 9). How does RNA polymerase transcribe DNA packaged in nucleosomes? Does the nucleosome have to be disassembled before the DNA within it can be transcribed? Surprisingly, RNA polymerase II is able to move past nucleosomes with the help of a protein complex called FACT (*f*acilitates *c*hromatin *t*ranscription), which removes histone H2A/H2B dimers from the nucleosomes leaving histone "hexasomes." After polymerase II moves past the nucleosome, FACT and other accessory proteins help redeposit the histone dimers, restoring nucleosome structure. Also, we should note that chromatin that contains genes actively being transcribed has a less compact structure than chromatin that contains inactive genes. Chromatin in which active genes are packaged tends to contain histones with lots of acetyl groups (Chapter 9), whereas chromatin with inactive genes contains histones with fewer acetyl groups. These differences are discussed further in Chapter 19.

TERMINATION BY CHAIN CLEAVAGE AND THE ADDITION OF 3′ POLY(A) TAILS

The 3′ ends of RNA transcripts synthesized by RNA polymerase II are produced by endonucleolytic cleavage of the primary transcripts rather than by the termination of transcription (■ **Figure 11.17**). The actual transcription termination events often occur at multiple sites that



■ **FIGURE 11.17** Poly(A) tails are added to the 3′ ends of transcripts by the enzyme poly(A) polymerase. The 3′-end substrates for poly(A) polymerase are produced by endonucleolytic cleavage of the transcript downstream from a polyadenylation signal, which has the consensus sequence AAUAAA.

are located 1000 to 2000 nucleotides downstream from the site that will become the 3′ end of the mature transcript. That is, transcription proceeds beyond the site that will become the 3′ terminus, and the distal segment is removed by endonucleolytic cleavage. The cleavage event that produces the 3′ end of a transcript usually occurs at a site 11 to 30 nucleotides downstream from a conserved polyadenylation signal, consensus AAUAAA, and upstream from a GU-rich sequence located near the end of the transcript. After cleavage, the enzyme **poly(A) polymerase** adds poly(A) tails, tracts of adenosine monophosphate residues about 200 nucleotides long, to the 3′ ends of the transcripts (Figure 11.17). The addition of poly(A) tails to eukaryotic mRNAs is called **polyadenylation.** To examine the polyadenylation signal of the human *HBB* (β-globin) gene, check out Solve It: Formation of the 3′-Terminus of an RNA Polymerase II Transcript.

The formation of poly(A) tails on transcripts requires a specificity component that recognizes and binds to the AAUAAA sequence, a stimulatory factor that binds to the GU-rich sequence, an endonuclease, and the poly(A) polymerase. These proteins form a multimeric complex that carries out both the cleavage and the polyadenylation in tightly coupled reactions. The poly(A) tails of eukaryotic mRNAs enhance their stability and play an important role in their transport from the nucleus to the cytoplasm.

In contrast to RNA polymerase II, both RNA polymerase I and III respond to discrete termination signals. RNA polymerase I terminates transcription in response to an 18-nucleotide-long sequence that is recognized by an associated terminator protein. RNA polymerase III responds to a termination signal that is similar to the rho-independent terminator in *E. coli* (see Figure 11.10).

RNA EDITING: ALTERING THE INFORMATION CONTENT OF mRNA MOLECULES

According to the central dogma of molecular biology, genetic information flows from DNA to RNA to protein during gene expression. Normally, the genetic information is not altered in the mRNA intermediary. However, the discovery of **RNA editing** has shown that exceptions do occur. RNA editing processes alter the information content of gene transcripts in two ways: (1) by changing the structures of individual bases and (2) by inserting or deleting uridine monophosphate residues.

The first type of RNA editing, which results in the substitution of one base for another base, is rare. This type of editing was discovered in studies of the apolipoprotein-B (*apo-B*) genes and mRNAs in rabbits and humans. Apolipoproteins are blood proteins that transport certain types of fat molecules in the circulatory system. In the liver, the *apo-B* mRNA encodes a large protein 4563 amino acids long. In the intestine, the *apo-B* mRNA directs the synthesis of a protein only 2153 amino acids long. Here, a C residue in the pre-mRNA is converted to a U, generating an internal UAA translation–termination codon, which results in the truncated apolipoprotein (■ **Figure 11.18**). UAA is one of three codons that terminates polypeptide chains during translation. If a UAA codon is produced within the coding region of an mRNA, it will prematurely terminate the polypeptide during translation, yielding an incomplete gene product. The C → U conversion is catalyzed by a sequence-specific RNA-binding protein with an activity that removes amino groups from cytosine residues. A similar example of RNA editing has been documented for an mRNA specifying a protein (the glutamate receptor) present in rat brain cells. More extensive mRNA editing of the C → U type occurs in the mitochondria of plants, where most of the gene transcripts are edited to some degree. Mitochondria have their own DNA genomes and protein-synthesizing

Solve It!

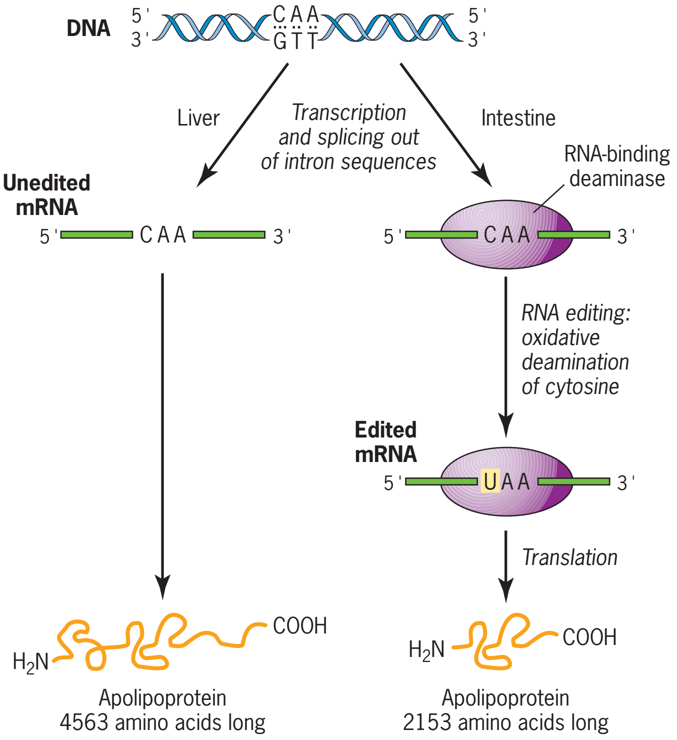
Formation of the 3′-Terminus of an RNA Polymerase II Transcript

The nucleotide sequence of the nontemplate strand of a portion of the human *HBB* (β-globin) gene and the carboxyl-terminus of its product, human β-globin (using the single-letter amino acid code; see Figure 12.1), are given as follows. Remember that the nontemplate strand will have the same sequence as the transcript of the gene, but with T's in place of U's.

```
5'–GGTGTGGCTA ATGCCCTGGC CCACAAGTAT CACTAAGCTC GCTTTCTTGC
     G  V  A   N  A  L  A   H  K  Y   H  COOH-terminus of β-globin
   TGTCCAATTT CTATTAAAGG TTCCTTTGTT CCCTAAGTCC AACTACTAAA
   CTGGGGGATA TTATGAAGGG CCTTGAGCAT CTGGATTCTG CCTAATAAAA
   AACATTTATT TTCATTGCAA TGATGTATTT AAATTATTTC TGAATATTT–3'
```

Note that every other codon is underlined in the coding region of the gene. Also, note that the GT-rich sequence involved in cleavage is located far downstream, near the end of the transcription unit, and is not shown. Can you predict the exact endonucleolytic cleavage site that produces the 3′ end of the transcript? Can you predict the approximate cleavage site? Will the 3′ end of the transcript produced by this cleavage event undergo any subsequent modification(s)? If so, what?

▶ ***To see the solution to this problem, visit the Student Companion site.***


DNA
5'
3'
CAA
GTT
5'
3'
Liver
Transcription and splicing out of intron sequences
Intestine
RNA-binding deaminase
Unedited mRNA
5' CAA 3'
5' CAA 3'
RNA editing: oxidative deamination of cytosine
Edited mRNA
5' UAA 3'
Translation
H_2N
COOH
H_2N
COOH
Apolipoprotein 4563 amino acids long
Apolipoprotein 2153 amino acids long

■ **FIGURE 11.18** Editing of the apolipoprotein-B mRNA in the intestines of mammals.

machinery (Chapter 15). In some transcripts present in plant mitochondria, most of the C's are converted to U residues.

A second, more complex type of RNA editing occurs in the mitochondria of trypanosomes (a group of flagellated protozoa that causes sleeping sickness in humans). In this case, uridine monophosphate residues are inserted into (occasionally deleted from) gene transcripts, causing major changes in the polypeptides specified by the mRNA molecules. This RNA editing process is mediated by **guide RNAs** transcribed from distinct mitochondrial genes. The guide RNAs contain sequences that are partially complementary to the pre-mRNAs to be edited. Pairing between the guide RNAs and the pre-mRNAs results in gaps with unpaired A residues in the guide RNAs. The guide RNAs serve as templates for editing, as U's are inserted in the gaps in pre-mRNA molecules opposite the A's in the guide RNAs.

Why do these RNA editing processes occur? Why are the final nucleotide sequences of these mRNAs not specified by the sequences of the mitochondrial genes as they are in most nuclear genes? As yet, answers to these interesting questions are purely speculative. Trypanosomes are primitive single-celled eukaryotes that diverged from other eukaryotes early in evolution. Some evolutionists have speculated that RNA editing was common in ancient cells, where many reactions are thought to have been catalyzed by RNA molecules instead of proteins. Another view is that RNA editing is a primitive mechanism for altering patterns of gene expression. For whatever reason, RNA editing plays a major role in the expression of genes in the mitochondria of trypanosomes and plants.

KEY POINTS

- *Three to five different RNA polymerases are present in eukaryotes, and each polymerase transcribes a distinct set of genes.*
- *Eukaryotic gene transcripts usually undergo three major modifications: (1) the addition of 7-methyl guanosine caps to 5′ termini, (2) the addition of poly(A) tails to 3′ ends, and (3) the excision of noncoding intron sequences.*
- *The information content of some eukaryotic transcripts is altered by RNA editing, which changes the nucleotide sequences of transcripts prior to their translation.*

Interrupted Genes in Eukaryotes: Exons and Introns

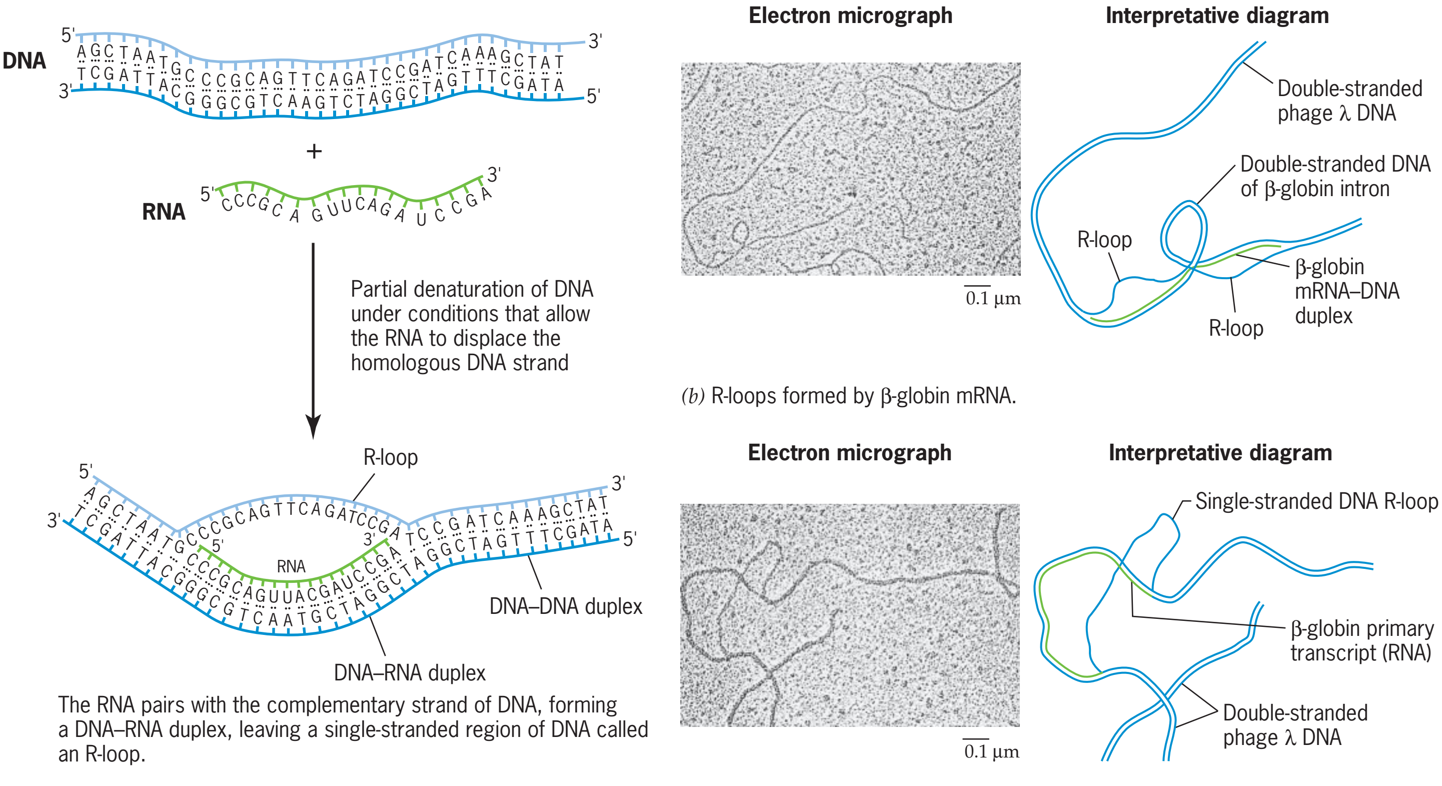
Most eukaryotic genes contain noncoding sequences called introns that interrupt the coding sequences, or exons. The introns are excised from RNA transcripts prior to their transport to the cytoplasm.

Most of the well-characterized genes of prokaryotes consist of continuous sequences of nucleotide pairs, which specify colinear sequences of amino acids in the polypeptide gene products. However, in 1977, molecular analyses of three eukaryotic genes yielded a major surprise. Studies of mouse and rabbit β-globin (one of two different proteins in hemoglobin) genes and the chicken ovalbumin (an egg storage protein) gene revealed that they contain noncoding sequences intervening between coding sequences. They were subsequently found in the nontranslated regions of some genes. They are called **introns** (for **int**ervening sequences.) The sequences that remain present in mature mRNA molecules (both coding and noncoding sequences) are called **exons** (for **ex**pressed sequences).

Some of the earliest evidence for introns in mammalian β-globin genes resulted from the visualization of genomic DNA–mRNA hybrids by electron microscopy. Because DNA–RNA duplexes are more stable than DNA double helices, when partially denatured DNA double helices are incubated with homologous RNA molecules under the appropriate conditions, the RNA strands will hybridize with the complementary DNA strands, displacing the equivalent DNA strands (■ **Figure 11.19*a***). The resulting DNA–RNA hybrid structures will contain single-stranded regions of DNA called **R-loops,** where RNA molecules have displaced DNA strands to form DNA–RNA duplex regions. These R-loops can be visualized directly by electron microscopy.

When Shirley Tilghman, Philip Leder, and colleagues hybridized purified mouse β-globin mRNA to a DNA molecule that contained the mouse β-globin gene, they observed two R-loops separated by a loop of double-stranded DNA (■ **Figure 11.19*b***). Their results demonstrated the presence of a sequence of nucleotide pairs in the middle of the β-globin gene that is not present in β-globin mRNA and, therefore, does not encode amino acids in the β-globin polypeptide. When Tilghman and coworkers repeated the R-loop experiments using purified β-globin gene transcripts isolated from nuclei and believed to be primary gene transcripts or pre-mRNA molecules, in place of cytoplasmic β-globin mRNA, they observed only one R-loop (■ **Figure 11.19*c***). This result indicated that the primary transcript contains the complete structural gene sequence, including both exons and introns. Together, the R-loop results obtained with cytoplasmic mRNA and nuclear pre-mRNA demonstrate that the intron sequence is excised and the exon sequences are spliced together during processing events that convert the primary transcript to the mature mRNA.

Tilghman and coworkers confirmed their interpretation of the R-loop results by comparing the sequence of the mouse β-globin gene with the predicted amino acid sequence of the β-globin polypeptide. Their results showed that the gene contained a noncoding intron at this position in the gene. Subsequent research showed that the mouse β-globin gene actually contains two introns. For details of these studies and additional information on the discovery of introns, see A Milestone in Genetics: Introns on the Student Companion site.


DNA
5'
AGCTAATGCCCGCAGTTCAGATCCGATCAAAGCTAT
3'
3'
TCGATTACGGGCGTCAAGTCTAGGCTAGTTTCGATA
5'
+
RNA
5'
CCCGCAGUUCAGAUCCGA
3'
R-loop
RNA
DNA–DNA duplex
DNA–RNA duplex
Electron micrograph
Interpretative diagram
0.1 μm
Double-stranded phage λ DNA
Double-stranded DNA of β-globin intron
R-loop
β-globin mRNA–DNA duplex
R-loop
Electron micrograph
Interpretative diagram
Single-stranded DNA R-loop
β-globin primary transcript (RNA)
Double-stranded phage λ DNA
0.1 μm

Partial denaturation of DNA under conditions that allow the RNA to displace the homologous DNA strand

The RNA pairs with the complementary strand of DNA, forming a DNA–RNA duplex, leaving a single-stranded region of DNA called an R-loop.

(*a*) The technique of R-loop hybridization.

(*b*) R-loops formed by β-globin mRNA.

(*c*) R-loop formed by β-globin primary transcript (pre-mRNA).

■ **FIGURE 11.19** R-loop evidence for an intron in the mouse β-globin gene. (*a*) R-loop hybridization. (*b*) When mouse β-globin genes and mRNAs were hybridized under R-loop conditions, two R-loops were observed in the resulting DNA–RNA hybrids. (*c*) When primary transcripts or pre-mRNAs of mouse β-globin genes were used in the R-loop experiments, a single R-loop was observed. These results demonstrate that the intron sequence is present in the primary transcript but is removed during the processing of the primary transcript to produce the mature mRNA.

SOME VERY LARGE EUKARYOTIC GENES

Subsequent to the pioneering studies on the mammalian globin genes and the chicken ovalbumin gene (see Milestone on the Student Companion site), noncoding introns have been demonstrated in a large number of eukaryotic genes. In fact, interrupted genes are much more common than uninterrupted genes in higher animals and plants. For example, the *Xenopus laevis* gene that encodes vitellogenin A2 (which ends up as egg yolk protein) contains 33 introns, and the chicken 1α2 collagen gene contains at least 50 introns. The collagen gene spans 37,000 nucleotide pairs but gives rise to an mRNA molecule only about 4600 nucleotides long. Other genes contain relatively few introns, but some of the introns are very large. For example, the *Ultrabithorax* (*Ubx*) gene of *Drosophila* contains an intron that is approximately 70,000 nucleotide pairs in length. The largest gene characterized to date is the human *DMD* gene, which causes Duchenne muscular dystrophy when rendered nonfunctional by mutation. The *DMD* gene spans 2.5 million nucleotide pairs and contains 78 introns.

Although introns are present in most genes of higher animals and plants, they are not essential because not all such genes contain introns. The sea urchin histone genes and four *Drosophila* heat-shock genes were among the first animal genes shown to lack introns. We now know that many genes of higher animals and plants lack introns.

INTRONS: BIOLOGICAL SIGNIFICANCE?

At present, scientists know relatively little about the biological significance of the exon–intron structure of eukaryotic genes. Introns are highly variable in size, ranging from about 50 nucleotide pairs to thousands of nucleotide pairs in length. This fact has led to speculation that introns may play a role in regulating gene expression. Although it is unclear how introns regulate gene expression, new research has shown that some introns contain sequences that can regulate gene expression in either a positive or negative fashion. Other introns contain alternative tissue-specific promoters; still others contain sequences that enhance the accumulation of transcripts. The fact that introns accumulate new mutations much more rapidly than exons indicates that many of the specific nucleotide-pair sequences of introns, excluding the ends, are not very important.

In some cases, the different exons of genes encode different functional domains of the protein gene products. This is most apparent in the case of the genes encoding heavy and light antibody chains (see Figure 20.17). In the case of the mammalian globin genes, the middle exon encodes the heme-binding domain of the protein. There has been considerable speculation that the exon–intron structure of eukaryotic genes has resulted from the evolution of new genes by the fusion of uninterrupted (single exon) ancestral genes. If this hypothesis is correct, introns may merely be relics of the evolutionary process.

Alternatively, introns may provide a selective advantage by increasing the rate at which coding sequences in different exons of a gene can reassort by recombination, thus speeding up the process of evolution. In some cases, alternate ways of splicing a transcript produce a family of related proteins. In these cases, introns result in multiple products from a single gene. The alternate splicing of the rat troponin T transcript is illustrated in Figure 19.2. In the case of the mitochondrial gene of yeast encoding cytochrome *b*, the introns contain exons of genes encoding enzymes involved in processing the primary transcript of the gene. Thus, different introns may indeed play different roles, and many introns may have no biological significance. Since many eukaryotic genes contain no introns, these noncoding regions are not required for normal gene expression.

KEY POINTS

- *Most, but not all, eukaryotic genes are split into coding sequences called exons and noncoding sequences called introns.*
- *Some genes contain very large introns; others harbor large numbers of small introns.*
- *The biological significance of introns is still open to debate.*

Removal of Intron Sequences by RNA Splicing

The noncoding introns are excised from gene transcripts by several different mechanisms.

Most nuclear genes that encode proteins in multicellular eukaryotes contain introns. Fewer, but still many, of the genes of unicellular eukaryotes such as the yeasts contain introns. Rare genes of archaea and of a few viruses of prokaryotes also contain introns. In the case of these "split" genes, the primary transcript contains the entire sequence of the gene, and the intron sequences are excised during RNA processing (see Figure 11.12).

For genes that encode proteins, the splicing mechanism must be precise; it must join exon sequences with accuracy to the single nucleotide to assure that codons in exons distal to introns are read correctly (■ **Figure 11.20**). Accuracy to this degree would seem to require precise splicing signals, presumably nucleotide sequences within introns and at the exon–intron junctions. However, in the primary transcripts of nuclear genes, the only completely conserved sequences of different introns are the dinucleotide sequences at the ends of introns, namely,

$$\text{exon-}\overbrace{\text{GT..............AG}}^{\text{intron}}\text{-exon}$$

The sequences shown here are for the DNA nontemplate strand (equivalent to the RNA transcript, but with T rather than U). In addition, there are short consensus sequences at the exon–intron junctions. For nuclear genes, the consensus junctions are

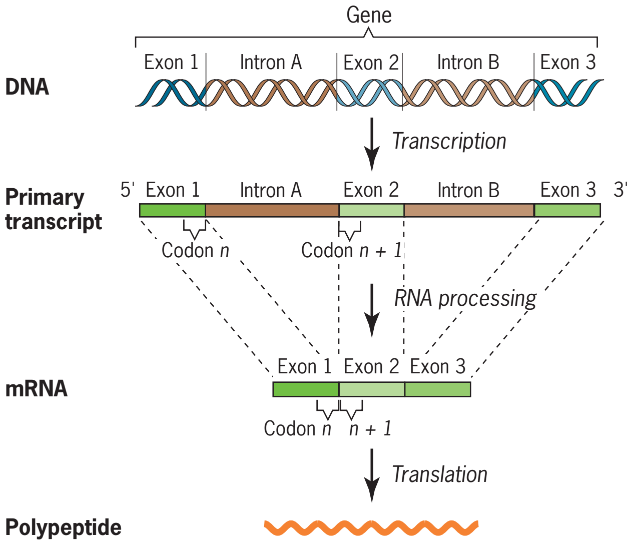
$$\overbrace{A_{64}G_{73}}^{\text{exon}}\overbrace{G_{100}T_{100}A_{68}A_{68}G_{84}T_{63}\text{............}6Py_{74\text{-}87}N\,C_{65}A_{100}G_{100}}^{\text{intron}}\;\overbrace{N}^{\text{exon}}$$

The numerical subscripts indicate the percentage frequencies of the consensus bases at each position; thus, a 100 subscript indicates that a base is always present at that position. N and Py indicate that any of the four standard nucleotides or either pyrimidine, respectively, may be present at the indicated position. The exon–intron junctions are different for tRNA genes and structural genes in mitochondria and chloroplasts, which utilize different RNA splicing mechanisms. However, different species do show some sequence conservation at exon–intron junctions.

Recent research has shown that splicing and intron sequences can influence gene expression. Direct evidence for their importance has been provided by mutations at these sites that cause mutant phenotypes in many different eukaryotes. Indeed, such mutations are sometimes responsible for inherited diseases in humans, such as hemoglobin disorders.

The discovery of noncoding introns in genes stimulated intense interest in the mechanism(s) by which intron sequences are removed during gene expression. The early demonstration that the intron sequences in eukaryotic genes were transcribed along with the exon sequences focused research on the processing of primary gene transcripts. Just as *in vitro* systems provided important information about the mechanisms of transcription and translation, the key to understanding RNA splicing events was the development of *in vitro* splicing systems. By using these systems, researchers have shown that there are three distinct types of intron excision from RNA transcripts.

1. The introns of tRNA precursors are excised by precise endonucleolytic cleavage and ligation reactions catalyzed by special splicing endonuclease and ligase activities.



■ **FIGURE 11.20** The excision of intron sequences from primary transcripts by RNA splicing. The splicing mechanism must be accurate to the single nucleotide to assure that codons in downstream exons are translated correctly to produce the right amino acid sequence in the polypeptide product.