10

GENE EXPRESSION

Transcription

STUDY OBJECTIVES

1. To examine the types of RNA and their roles in gene expression 245, 256
2. To look at the process of transcription, including start and stop signals, in both prokaryotes and eukaryotes 246
3. To investigate posttranscriptional changes in eukaryotic messenger RNAs, including an analysis of intron removal 260

STUDY OUTLINE

A computer model of the serine transfer RNA. The amino acid binding site is *yellow;* the anticodon is *red.* (© Ken Eward/SPL/Photo Researchers.)

In this chapter, we continue our study of genetics at the molecular level. After discussing the structure of DNA and the way in which it replicates in the last chapter, we turn our attention here and in the next chapter to the way in which the genetic material—primarily DNA—expresses itself. In this chapter, we concentrate on the conversion of DNA information into RNA information, the first step in gene expression. In the next chapter, we look at the conversion of RNA information into proteins. Later chapters discuss the control of these processes. We begin with prokaryotes and later in the chapter discuss the conceptually similar but functionally more complex process in eukaryotes.

All living things synthesize proteins. In fact, the types of proteins that a cell synthesizes determine the kind of cell it is. Hence, the genetic material must determine the types and quantities of proteins a cell synthesizes. Proteins (polypeptides) are made up of strings of amino acids (three hundred to five hundred, on average) joined together by peptide bonds. (We cover protein structure and synthesis in chapter 11.) Each protein contains a unique combination of only twenty amino acids. The amino acid sequence is specified by the sequence of nucleotides in DNA or RNA. In all prokaryotes, eukaryotes, and DNA viruses, the gene is a sequence of nucleotides in DNA that codes for the sequence of RNA. That RNA then determines which amino acids are included in a polypeptide. RNA usually serves as an intermediary between DNA and proteins. (In RNA viruses, the RNA may serve as a template for the eventual synthesis of DNA, or the RNA may serve as genetic material without DNA ever being formed. We will consider these cases at the end of the chapter.)

In 1958, Francis Crick originally described the flow of genetic information as the **central dogma:** DNA transfers information to RNA, which then directly controls protein synthesis (fig. 10.1). DNA also controls its own replication. **Transcription** is the process of synthesizing RNA from a DNA template using the rules of complementarity—the DNA information is rewritten, but in the same nucleotide language. RNA controls the synthesis of proteins in a process called **translation** because the information in the language of nucleotides is *translated* into information in the language of amino acids.

In the previous chapter, we introduced the idea of proteins that recognize specific DNA sequences and bind to those sequences. Specifically, we introduced the initiator proteins that bind to *oriC* and the proteins that bind to the terminator sequences. DNA polymerases and some of the other proteins involved in DNA replication bind to DNA, but they do not necessarily bind to any specific sequences. Proteins that recognize specific DNA sequences are critically important to the transcriptional process. In the next chapter, we spend more time on proteins, discussing their structures and how they are synthesized. It is sufficient to say here that specific proteins recognize

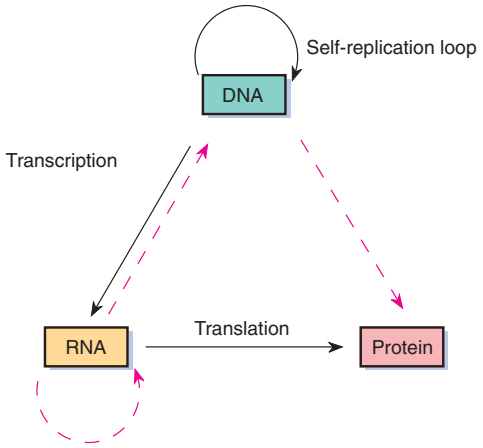
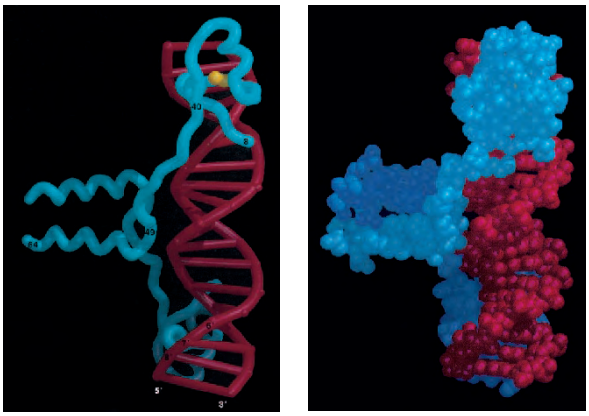


Figure 10.1 Crick's original central dogma depicted the flow of genetic information. Dashed *red* lines indicate the possible information transfers unconfirmed in 1958, when Crick proposed the central dogma.

specific DNA sequences. They do so by interdigitating the amino acid side chains of the proteins into the grooves of the DNA, thereby recognizing specific sequences by hydrogen bonding and other electrostatic interactions between the side chains of the amino acids of the proteins and the bases of the DNA (fig. 10.2). Proteins can have parts that recognize DNA sequences and parts that recognize other proteins or that perform other enzymatic activities such as hydrolyzing ATP.



(a) (b)

Figure 10.2 Computer model of the interaction of a yeast transcriptional factor, GAL4 (*blue*), and a seventeen-base-pair region of DNA (*red*). Zinc ions are in *yellow.* The protein is a dimer; only the DNA recognition region and associated part are shown. Part (*b*) is a space-filling model of part (*a*). (Reprinted with permission from *Nature,* 2 April 1992, Vol. 356, p. 411, fig. 3b,c. Copyright 1992 Macmillan Magazines Limited.)

TYPES OF RNA

In the protein synthesis process, three different kinds of RNA serve in three different roles. The first type is **messenger RNA (mRNA),** which carries the DNA sequence information to particles in the cytoplasm known as **ribosomes,** where the messenger RNA is translated. The second type is **transfer RNA (tRNA),** which brings the amino acids to the ribosomes, where protein synthesis takes place. The third type of RNA is a structural and functional part of the ribosome called **ribosomal RNA (rRNA).** The general relationship of the roles of these three types of RNA is diagrammed in figure 10.3. In addition, small RNAs play other roles in cellular metabolism, some of which are described later in the chapter.

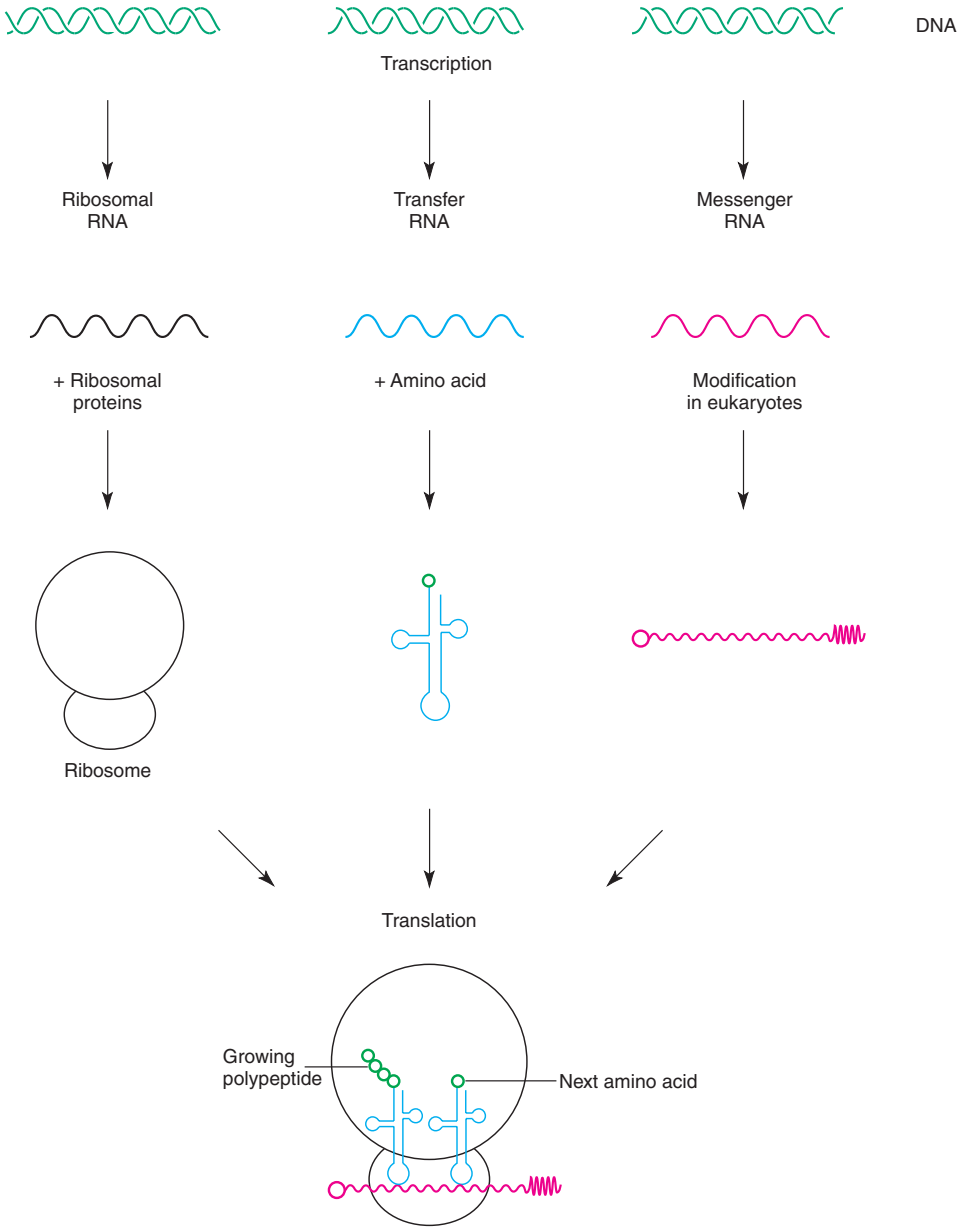


Figure 10.3 Relationship among the three types of RNA—ribosomal, transfer, and messenger—during protein synthesis. All three types are found together at the ribosome during protein synthesis.

We know that DNA does not take part directly in protein synthesis because, in eukaryotes, translation occurs in the cytoplasm, whereas DNA remains in the nucleus. We suspected for a long time that the genetic intermediate in prokaryotes and eukaryotes was RNA because the cytoplasmic RNA concentration increases with increasing protein synthesis, and the cytoplasmic RNAs carry nucleotide sequences complementary to the cell's DNA. Proof of an RNA intermediate came when it was shown that messenger RNA directs protein synthesis.

PROKARYOTIC DNA TRANSCRIPTION

DNA-RNA Complementarity

What proof do we have that a messenger RNA exists? That is, what proof convinced geneticists that gene-sized RNAs (not transfer RNAs or ribosomal RNAs) were found in the cytoplasm that were complementary to the DNA in the nucleus? At least two lines of evidence exist. First, it was shown that the RNAs produced by various organisms have base ratios very similar to the base ratios in the same organisms' DNA (table 10.1). The second line of evidence comes from experiments by B. Hall, S. Spiegelman, and others using **DNA-RNA hybridization.** This technique denatures DNA by heating, which causes the two strands of the double helix to separate. When the solution cools, a certain proportion of the DNA strands rejoin and rewind—that is, complementary strands "find" each other and re-form double helices. When RNA is added to the denatured DNA solution and the solution is cooled slowly, some of the RNA forms double helices with the DNA if the RNA fragments are complementary to a section of the DNA (fig. 10.4). The existence of extensive complementarity between DNA and RNA is a persuasive indication that DNA acts as a template for complementary RNA.

In another experiment, DNA-RNA hybridization showed that bacteriophage infection led to the production of phage-specific messenger RNA. Gene-sized pieces of RNA extracted from *Escherichia coli* before and after bacteriophage T2 infection were tested to see if they hybridized with the DNA of the T2 phage or with the DNA of the *E. coli* cell. The RNA in the *E. coli* cell was found to hybridize with the *E. coli* DNA before infection but with the T2 DNA after infection. Thus it is apparent that when the phage attacks the *E. coli* cell, it starts to manufacture RNA complementary to its own DNA and stops the *E. coli* DNA from serving as a template.

Having reached the conclusion that RNA is transcribed (synthesized) from a DNA template and then directs protein synthesis, we look at two questions. First, is this RNA single- or double-stranded? Second, is it synthesized (transcribed) from one or both strands of the parental DNA?

Table 10.1 **Correspondence of Base Ratios Between DNA and RNA of the Same Species**

	RNA % G + C	DNA % G + C
E. coli	52	51
T2 phage	35	35
Calf thymus gland	40	43

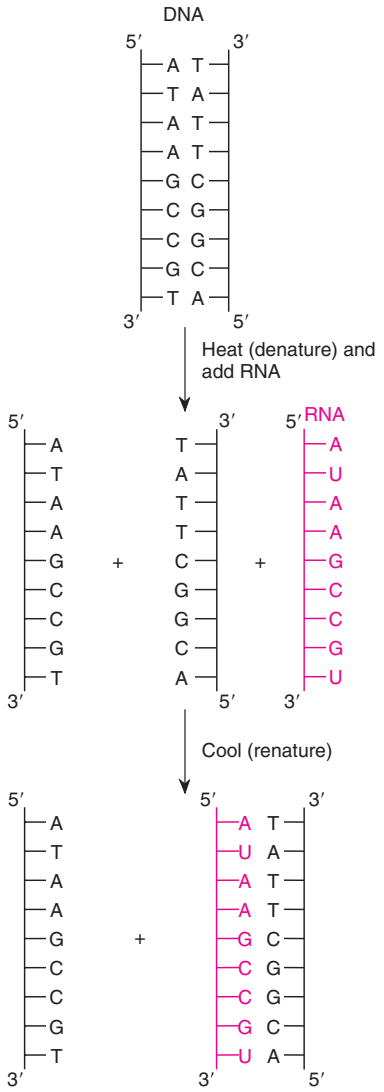


Figure 10.4 DNA-RNA hybridization occurs between DNA and complementary RNA.

For the most part, cellular RNA does not exist as a double helix. It can form double helical sections when complementary parts come into apposition (e.g., see fig. 10.16), but its general form is not a double helix. The simplest, and most convincing, evidence for this is that complementary RNA bases do not occur in corresponding proportions (Chargaff's ratios). That is, in RNA, uracil does not usually occur in the same quantity as adenine, nor does cytosine occur in the same quantity as guanine (table 10.2).

The answer to the second question is that RNA is not usually copied from both strands of any given segment of the DNA double helix, although rare exceptions do occur. Consider a sequence of nucleotides on one strand of a DNA duplex that specifies a sequence of amino acids for a protein, with the complementary nucleotide sequence also specifying the amino acid sequence for another functional protein. Since most enzymes are three hundred to five hundred amino acids long, the virtual impossibility of this task is obvious. It was, therefore, assumed *a priori* that, for any particular gene—that is, in any particular segment of DNA—the sequence on only one strand is transcribed and its complementary sequence is not. There is now considerable evidence to support this assumption.

The most impressive evidence that only one DNA strand transcribes RNA comes from work done with bacteriophage SP8, which attacks *Bacillus subtilis.* This phage has an interesting property—a great disparity in the purine-pyrimidine ratio of the two strands of its DNA. The disparity is significant enough that the two strands can be separated by density using density-gradient centrifugation. After denaturation and separation of the two strands, DNA-RNA hybridization can be carried out separately on each of the two strands with the RNA produced after the virus infects the bacterium. J. Marmur and his colleagues found that hybridization occurred only between the RNA and the heavier of the two DNA strands. Thus, only the heavy strand acted as a template for the production of RNA during the infection process.

The idea that only one strand of DNA serves as a transcription template for RNA has also been verified for several other small phages. However, when we get to larger viruses and cells, we find that either of the strands may be transcribed, but only one strand is used as a template in any one region. This was clearly shown in phage T4 of *E. coli,* where certain RNAs hybridize with one DNA strand, and other RNAs hybridize with the other. Let us now look at the transcription process in prokaryotes, then proceed to examine the three types of RNA in detail, and finally look at transcription in eukaryotes.

Table 10.2 Base Composition in RNA (percentage)

	Adenine	Uracil	Guanine	Cytosine
E. coli	24	22	32	22
Euglena	26	19	31	24
Poliovirus	30	25	25	20

Prokaryotic RNA Polymerase

In prokaryotes, transcription of RNA is controlled by **RNA polymerase.** Using DNA as a template, this enzyme polymerizes ribonucleoside triphosphates (RNA nucleotides). The complete RNA polymerase enzyme of *E. coli*—the holoenzyme—is composed of a core enzyme and a **sigma factor.** The core enzyme is composed of four subunits: α (two copies), β, and β'; this core is the component of the holoenzyme that actually carries out polymerization. The sigma factor is involved in recognizing transcription start signals on the DNA. Following the initiation of transcription, the sigma factor disassociates from the core enzyme.

Logically, transcription should not be a continuous process like DNA replication. If there were no control of protein synthesis, all the cells of a higher organism would be identical, and a bacterial cell would be producing all of its proteins all of the time. Since some enzymes depend on substrates not present all of the time, and since some reactions in a cell occur less frequently than others, the cell—be it a bacterium or a human liver cell—needs to regulate its protein synthesis. One of the most efficient ways for a cell to exert the necessary control over protein synthesis is to perform transcription selectively. Transcription of nongenic regions or of genes coding for unneeded enzymes is wasteful. Therefore, RNA polymerase should be selective. It should use as transcription templates only those DNA segments (genes or small groups of genes) whose products the cell needs at that particular time.

The mechanisms of transcriptional control need to be examined in two ways. First, we need to understand how the beginnings and ends of transcribable sections (a single gene or a series of adjacent genes) are demarcated. Second, we need to understand how the cell can selectively repress or enhance transcription of certain of these transcribable sections. The latter issues—the keys to bacterial efficiency and eukaryotic growth and development—are covered in chapters 14 and 16, respectively.

RNA polymerase must be able to recognize both the beginnings and the ends of genes (or gene groups) on the DNA double helix in order to initiate and terminate transcription. It must also be able to recognize the correct DNA strand to avoid transcribing the DNA strand that is not informational. RNA polymerase accomplishes those tasks by recognizing certain start and stop signals in DNA, called initiation and termination sequences, respectively.

Prokaryotic Initiation and Termination Signals for Transcription

The DNA region that RNA polymerase associates with immediately before beginning transcription is known as the **promoter.** The promoter is an important part of gene expression in both prokaryotes and eukaryotes. Promoters contain the information for transcription initiation and are the major sites in which gene expression is controlled.

Without the sigma factor, the core enzyme of RNA polymerase binds randomly along the DNA. Formation of the holoenzyme brings about high affinity of RNA polymerase for DNA sequences in the promoter region. Termination of transcription comes about when the polymerase enzyme recognizes a DNA region known as a **terminator sequence.** Let us elaborate on the various stages of transcription (in this section and in boxes 10.1 and 10.2).

Promoters

The RNA polymerase molecule covers a region of about sixty base pairs of DNA. This was determined by causing the polymerase to bind to DNA and then digesting the mixture with nucleases, in a technique known as **footprinting** (fig. 10.5). The polymerase "protects" or prevents degradation of the region it covers. The undigested DNA is then isolated and its size determined. Geneticists have gained much new information about the nature of recognition regions within promoters through recombinant DNA technology and nucleotide sequencing techniques (see chapter 13). Sequencing of numerous promoters has shown that they contain common sequences. If the promoter nucleotide sequences align with each other, and each has exactly the same series of nucleotides in a given segment, we say that the sequence of that segment comprises an invariant or **conserved sequence.** If, however, there is some variation in the sequence, but certain nucleotides occur at a high frequency (significantly greater than by chance), we refer to those nucleotides as making up a **consensus sequence.** Surrounding a point in prokaryotic promoters about ten nucleotides before the first transcribed base is just such a consensus sequence—TATAAT. This sequence is known as a **Pribnow box** after one of its discoverers (fig. 10.6).

The nucleotides in the Pribnow box are mostly adenines and thymines, so the region is primarily held together by only two hydrogen bonds per base pair. Since local DNA denaturation occurs during transcription by RNA polymerase (the DNA is opened to allow transcription), fewer hydrogen bonds make this process easier energetically. When the polymerase is bound at the promoter region (fig. 10.6), it is in position to begin polymerization six to eight nucleotides down from the Pribnow box.

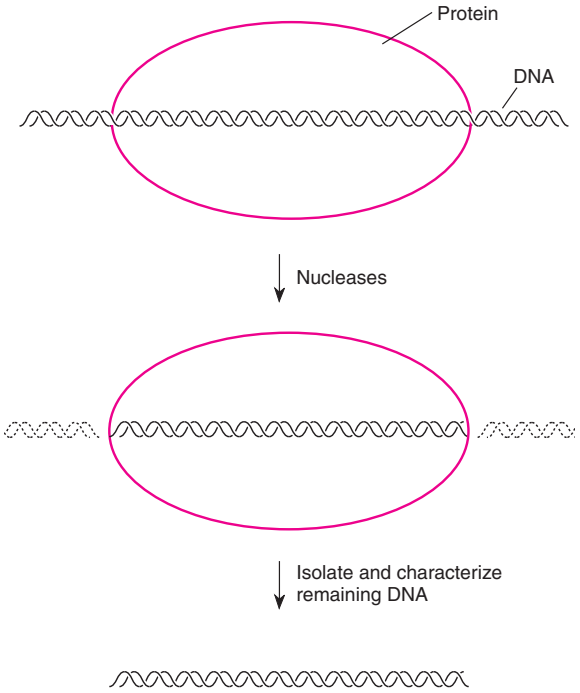


Figure 10.5 Footprinting technique. DNA in contact with a protein (e.g., RNA polymerase) is protected from nuclease degradation. The protected DNA is then isolated and characterized.

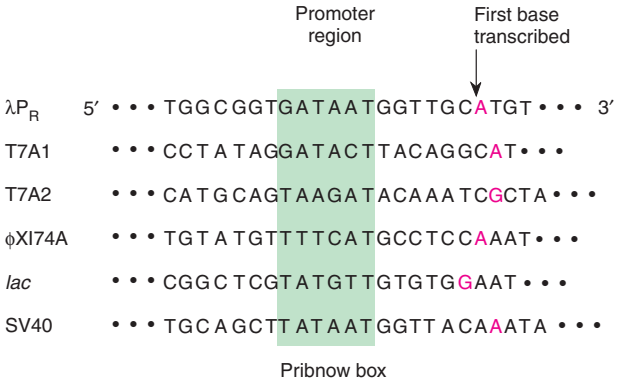


Figure 10.6 Nucleotide sequences of the promoter region and the first base transcribed from several different genes. Lambda (λ), T7, and ϕ×174 are bacteriophages. *Lac* is an *E. coli* gene, and SV40 is an animal virus. Only the SV40 promoter has the actual consensus sequence of TATAAT. Even when other sequenced promoters not shown here are Included, no base is found 100% of the time (conserved).

The sequences shown in figure 10.6 are those of the **coding strand** of DNA. It is a general convention to show the coding strand because both that strand and messenger RNA have the same sequences, substituting U for T in RNA; they are both complementary to the same **template strand** (also referred to as the **anticoding strand** or **noncoding strand;** fig. 10.7). Another convention is to indicate the first base transcribed by number +1 and to use positive numbers to count farther down the DNA in the **downstream** direction of transcription. If transcription is proceeding to the right, the direction to the left is called **upstream,** with bases indicated by negative numbers (fig. 10.8). Under this convention, the Pribnow box is often referred to as the −10 sequence.

Figure 10.8 also indicates another region with similar sequences among many promoters centered near −35 and referred to as the −35 sequence. The consensus sequence at −35 is TTGTCA. Mutation studies have attempted to determine the relative roles of the −10 and −35 sequences in transcription. In other words, mutations of bases in the −10 and −35 regions were examined to determine how they affected transcription initiation. The conclusions from these studies are that both regions contribute to the efficiency of polymerase binding. In other words, the more each sequence differs from the consensus sequence, the less frequently that promoter initiates transcription. The sigma factor recognizes both the −35 and the −10 sequences. The sigma factor is also sensitive to the spacing between these sequences, preferring (being most efficient at) seventeen base pairs.

Farther upstream from the −35 sequence is a recognition element in bacterial promoters that are very strongly expressed, such as the ribosomal RNA genes (fig. 10.8). This **upstream element,** or **UP element,** is about twenty base pairs long, is centered at −50, and is rich in A and T. By mutational studies, it has been shown that adding this element to promoters that don't normally have it greatly increases the rate of transcription. There are other recognition sites in prokaryotes, both upstream and downstream, at which various proteins attach that can enhance or inhibit transcription by direct contact with the polymerase (the α and σ subunits). We discuss these in chapter 14 under control of transcription. They are not part of what we think of as the core promoter, the DNA sequence needed for efficient binding of RNA polymerase.

Since the holoenzyme recognizes consensus sequences in a promoter, it is not surprising that some promoters are bound more efficiently than others or that different sigma factors exist within a cell. In *E. coli,* the major sigma factor is a protein of 70,000 daltons, referred to as σ^{70}. (One dalton is an atomic mass of 1.0000, approximately equal to the mass of a hydrogen atom.) The existence of about five less common sigma factors provides the cell with a mechanism for transcribing different genes under different circumstances. For example, in an *E. coli* cell subjected to elevated temperatures, a group of new proteins, referred to as **heat shock proteins,** appear, acting to protect the cell to some extent against the elevated temperatures. These proteins all appear at once because they have promoters that a different sigma factor recognizes, one with a molecular weight of 32,000 daltons (σ^{32}); this new sigma factor is produced by the cell after

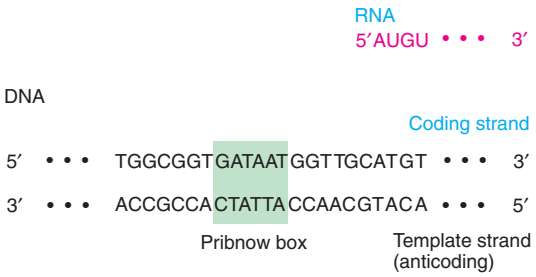


Figure 10.7 The template (anticoding) strand of DNA is complementary to both the coding strand and the transcribed RNA. The sequences are from the promoter of the λP_R region (see fig. 10.6).

Figure 10.8 Promoter of the *Escherichia coli* ribosomal RNA gene, *rrnB.* Note the −10 and −35 sequences and the upstream element. The first base transcribed (the transcriptional start site) is noted (+1), as well as the upstream, downstream, and transcription directions. (Data from W. Ross, et al., 1993. *Science* 262:1407.)

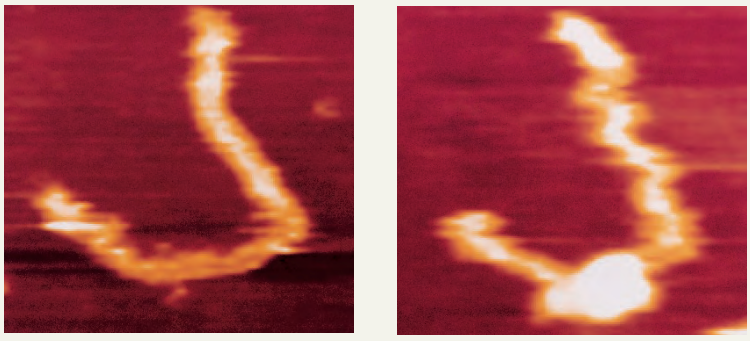
BOX 10.1

Experimental Methods

Observing Transcription in Real Time

The overwhelming evidence that molecular events, such as transcription, take place comes from genetic and biochemical analyses and occasionally an electron micrograph of one type or another (fig. 1). Thus, it is refreshing and illuminating to be able to observe some of the processes we know are taking place in real time; that is, to sit at a microscope and actually see these events happen. Such a study on transcription was published in 1991 in *Nature* by four scientists at Washington University in St. Louis.

Although new methods of microscopy are being developed, normally we cannot see these molecular events taking place; the components are too small. Making them visible in electron microscopes usually requires fixation that destroys the ability of the components to actually continue their tasks. The Washington University group overcame this by attaching a gold particle to DNA, thus rendering the motion of that DNA visible under the light microscope (fig. 2). The scientists immobilized the RNA polymerase to a glass coverslip; thus, as transcription took place, the DNA moved and the length of the tether of the gold particle increased. At first they stopped the process by limiting the concentration of nucleoside triphosphates (NTPs). They



(a) (b)

Figure 1 Visualizing transcription. Image of DNA before (*a*) and after (*b*) *E. coli* RNA polymerase (bright oblong object in *b*) binds to a promoter. Pictures are by scanning force microscopy, a new laser technique that images molecules in water. Image sizes are 300 by 300 nm. *Dark brown* represents substrate level; the highest point is *white* at about 10 nm high. Intermediate colors represent intermediate heights. (Courtesy of Martin Guthold and Carlos Bustamante, Institute of Molecular Biology and HHMI, University of Oregon.)

heat shock. We discuss heat shock proteins and other systems of transcriptional control in chapters 14 and 16.

From mutational studies of promoters and the proteins in the RNA polymerase holoenzyme, we now have a picture of a holoenzyme that sets down on a DNA promoter because the sigma factor recognizes the −10 and −35 elements, the α proteins recognize the UP element, and the α and σ subunits recognize proteins bound to various other upstream elements, when present (fig. 10.9*a*). This initiation complex is initially referred to as a closed complex because the DNA has not melted, which is the next step in transcription initiation (fig. 10.9*b*). After the transcription of 5–10 bases, the sigma factor is released (fig. 10.9*c* and *d*).

About seventeen base pairs of DNA are opened, and as transcription proceeds, about twelve bases of RNA form a DNA-RNA duplex at the point of transcription. Some of this information comes from studies with potassium permanganate ($KMnO_4$), which modifies DNA bases that are single-stranded but not double-stranded. Thus, the lengths of melted DNA can be determined experimentally. Also used is the technique of **photocrosslinking,** in which two moieties such as DNA and one or two proteins are caused to be permanently crosslinked, verifying their close contact. This is done by attaching a chemical crosslinking element to one of the moieties and then causing crosslinking to occur by shining light, usually ultraviolet, on the mixture.

Transcription, like DNA replication, always proceeds in the 5′ → 3′ direction. That is, a single base is added *de novo* and then new RNA nucleotides are added to the 3′-OH free end, as in DNA replication. However, unlike

could then observe the motion of the gold ball when no transcription was taking place. The scientists predicted that an immobilized gold ball would not move, and a tethered gold ball would show a limited amount of Brownian motion. That is, it would show a limited amount of blur in light microscope video images averaged over time. However, as soon as NTPs were added, any tethered gold ball would show an increased blur as it moved out of the field of vision and eventually would be released when transcription was completed. That is exactly what they saw (fig. 3). Thus, they succeeded in watching transcription take place in real time.

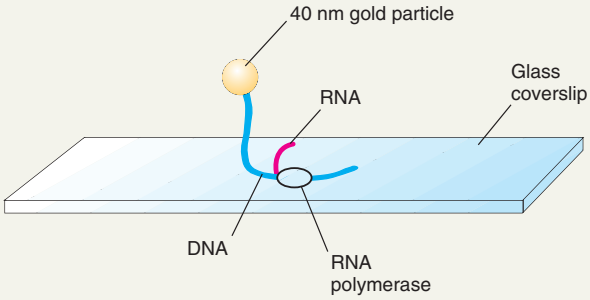


Figure 2 An experimental design in visualizing transcription in real time under the light microscope. Here, an RNA polymerase is immobilized on a coverslip, waiting for nucleoside triphosphates (NTPs) to be added. The gold particle is tethered by the DNA, allowing us to keep track visually of the end of the DNA.

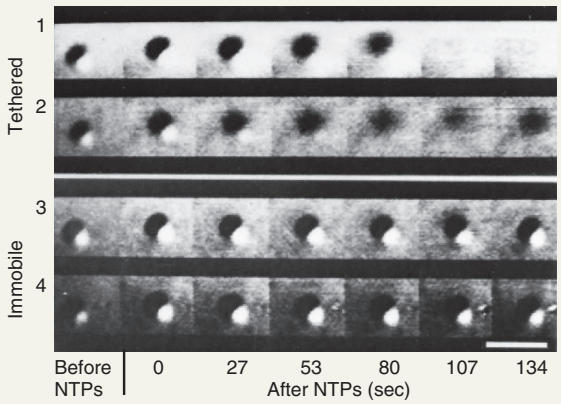


Figure 3 Enhanced light microscope images of the gold particles. In *3* and *4,* presumably immobilized particles show no change in focus over time. In *1* and *2,* Brownian motion—and hence blur—increases through time consistent with the lengthening of the tether (transcription). The particle in *1* was released 87 seconds (and in *2,* 135 seconds) after NTPs were added. Scale bar is 1 μm. (Transcription by single molecules of RNA polymerase observed by light microscopy. Robert Landick, Department of Biology, Washington University, St. Louis, MO.)

DNA polymerase, prokaryotic RNA polymerase does not seem to proofread as it proceeds. That is, RNA polymerase evidently does not verify the complementarity of the new bases added to the growing RNA strand. This deficiency is not serious; since many messenger RNAs are short-lived and many copies are made from actively transcribed genes, an occasional mistake will probably not produce permanent or overwhelming damage. If a particular RNA is not functional, a new one will be made soon. Evolutionarily speaking, it seems that it is more important to make RNA quickly than to proofread each RNA made.

Terminators

Transcription continues as RNA polymerase adds nucleotides to the growing RNA strand according to the rules of complementarity (C, G, A, and U of RNA pairing with G, C, T, and A of DNA, respectively). The polymerase moves down the DNA until the RNA polymerase reaches a stop signal, or terminator sequence. Two types of terminators, rho-dependent and rho-independent, differ in their dependency on the **rho protein** (Greek letter ρ). The functional form of rho is a hexamer, six identical copies of the protein. **Rho-independent terminators** cause termination of transcription even if rho is not present. **Rho-dependent terminators** require the rho protein; without it, RNA polymerase continues to transcribe past the terminator in a process known as **read-through.** Both types of terminators sequenced so far have one thing in common: They include a sequence and its inverted form separated by another short sequence, all together forming an **inverted-repeat sequence.** The

BOX 10.2

Experimental Methods

Polymerase Collisions: What Can a Cell Do?

Both RNA polymerase and DNA polymerase move along the DNA of a cell during a cell cycle. The DNA polymerase moves along at about ten times the speed of the transcribing enzyme. Since many genes are usually active in a cell, the interaction (collision) of the two enzymes is inevitable. What happens when this collision takes place? What does the cell do? Although we cannot directly observe these interactions, various bits of data suggest that a head-on collision could be fatal to the cell, and certain patterns of gene placement minimize the chance of a head-on collision.

B. Brewer first analyzed the problem of the coexistence of these two enzymes in a paper published in 1988 in the journal *Cell.* In evolutionary terms, the cell could obviate the problems of a head-on collision by either avoiding them or resolving them. Resolution would entail some sort of right-of-way settlement when the two enzymes met; for example, the RNA polymerase could drop off the DNA when a confrontation takes place. The cell might avoid confrontations if the genes are oriented so that transcription occurs for the most part in the same direction as DNA replication. That is, DNA replication begins at *oriC,* with Y-junctions proceeding to the left and the right until they meet 180 degrees later. Thus, to avoid head-on collisions, genes on the left and right arcs of the bacterial chromosome could be transcribed away from the origin of replication (fig. 1).

Brewer analyzed the orientation of genes on the *E. coli* chromosome; more recently, D. Zeigler and D. Dean did the same for the chromosome of *Bacillus subtilis.* In *B. subtilis,* 95% (91 of 96) of the genes analyzed were in the proper orientation to avoid a head-on collision of polymerases. Among the exceptions were sporulation genes, genes that would not be transcribed during DNA synthesis and whose orientation is thus not relevant to DNA polymerase activity. In *E. coli,* Brewer found that, overall, 74% (375 of 501) of the genes she looked at were oriented to avoid head-on collisions. Brewer's data were more impressive when she broke them down according to transcription function and activity.

For genes that transcribe very actively most of the time, the orientation is about 90% in the "safe" direction. For regulatory genes that are transcribed only very rarely, the orientation is random (50% safe). For other genes, the orientation was 72% in the safe direction. Thus, an organization clearly exists within the bacterial chromosome that helps to avoid head-on collisions of the two polymerases.

Brewer also provided evidence that a head-on collision between polymerases could be fatal to the cell. Studies selected inversions of the *E. coli* chromosome to see the effects of collision. (Inversions are regions that have been cut out and put back in the opposite orientation.) It was impossible to isolate inversion mutations that changed the orientation of genes in respect to *oriC.* Thus, it appears that a cell may not be able to resolve a head-on collision of polymerases and that evolution has solved the problem by having gene transcription generally oriented in the same direction as DNA replication.

More amazingly, Alberts and his colleagues recently studied what happens when a replication fork catches up to a stalled RNA polymerase. Not only does the replication fork pass the transcription apparatus, but the RNA polymerase can resume transcription after the replication fork passes without loss of the transcript. Although there are contrary observations in other systems, it appears that gene orientation and the behavior of polymerases allow cells to survive with both replication and transcription occurring on the same DNA.

terminator in figure 10.10 has the sequence AAAGGCTCC, 5′ → 3′, from both the left on the coding strand and from the right on the template strand. A four-base-pair sequence separates the inverted repeats. Inverted repeats can form a **stem-loop structure** by pairing complementary bases within the transcribed messenger RNA.

Both rho-dependent and rho-independent terminators have the stem-loop structure in RNA just before the last base transcribed. Rho-independent terminators, as figure 10.10 shows, also have a sequence of thymine-containing nucleotides after the inverted repeat, whereas rho-dependent terminators do not. Although the exact sequence of events at the terminator is not fully known, it appears that the RNA stem-loop structure forms and causes the RNA polymerase to pause just after completing it. This pause may then allow termination under two different circumstances.

In rho-independent terminators, the pause may occur just after the sequence of uracils is transcribed (fig. 10.11). Uracil-adenine base pairs have two hydrogen bonds and are thus less stable thermodynamically than guanine-cytosine base pairs. Perhaps during the pause, the uracil-adenine base pairs spontaneously denature, releasing the transcribed RNA and the RNA polymerase,

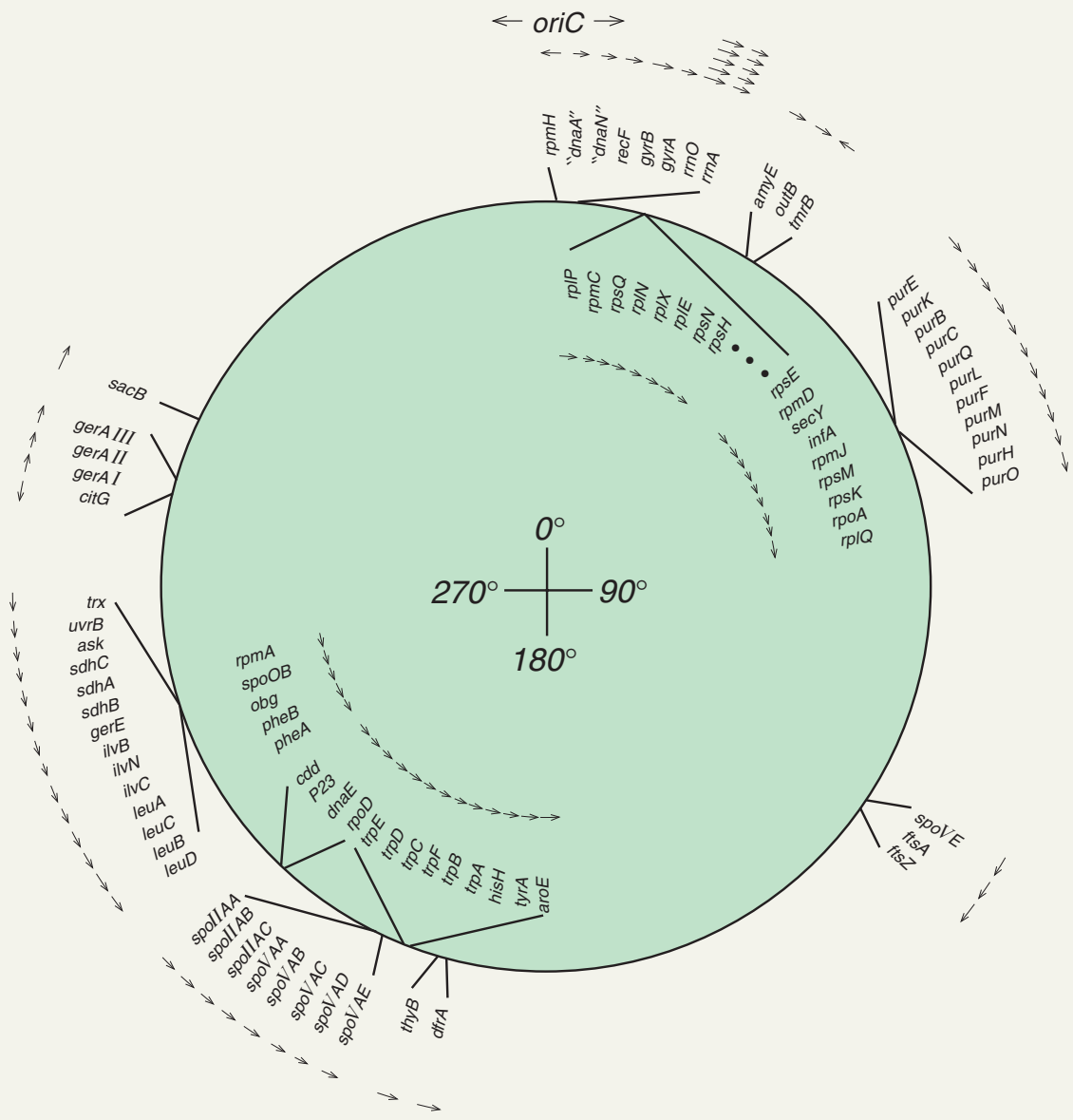


Figure 1 Location and orientation of gene transcription on the chromosome of *Bacillus subtilis* (*arrows*). DNA replication begins at *oriC* and terminates approximately 180 degrees from the origin of replication. Note that the overwhelming number of arrows point away from the origin of replication toward the termination point. (From D. R. Zeigler and D. H. Dean, "Orientation of genes in the *Bacillus subtilis* chromosome," *Genetics,* 125:703–8. Copyright © 1990 Genetics Society of America.)

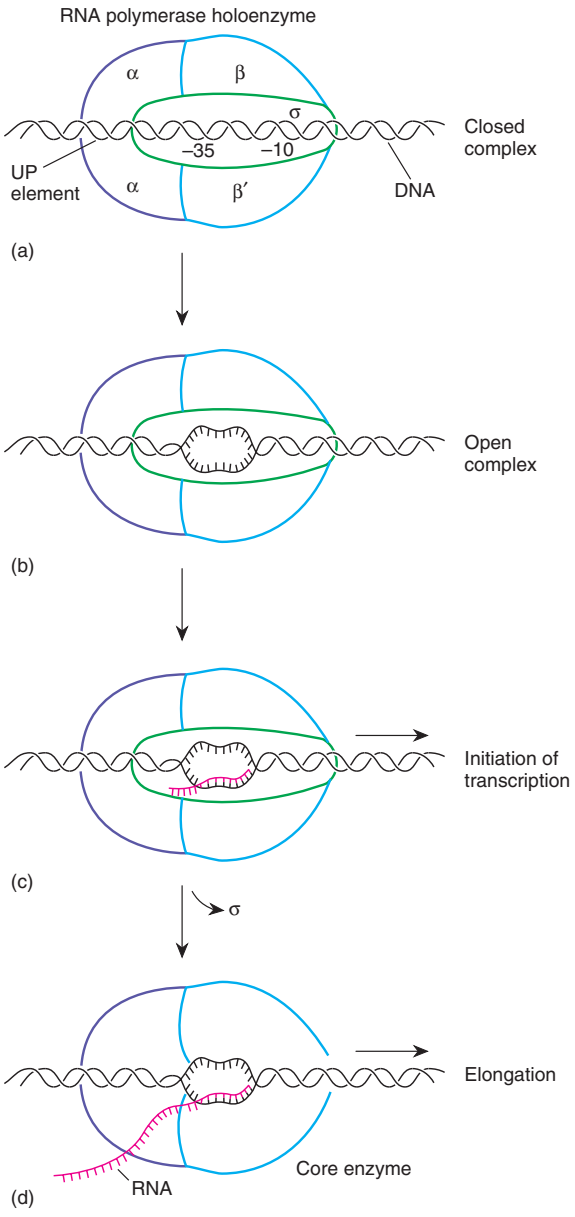


Figure 10.9 Transcription begins after RNA polymerase attaches to the promoter, with specificity imparted by the sigma factor. The DNA opens to form the open complex, transcription begins, the sigma factor leaves, and elongation commences.

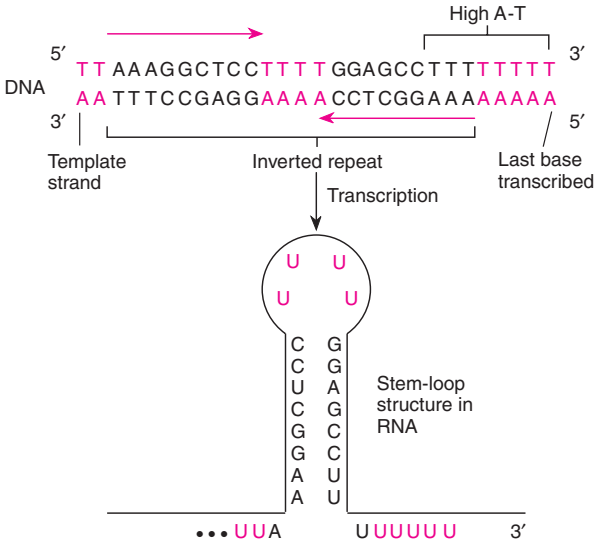


Figure 10.10 An inverted-repeat base sequence characterizes terminator regions of DNA. Stem-loop structures can occur as the RNA forms because of complementary sequences. The 3′ poly-U tail indicates a rho-independent terminator.

terminating the process, and making the polymerase available for further transcription of other promoters.

Rho-dependent terminators do not have the uracil sequence after the stem-loop structure. Here, termination depends on the action of rho, which appears to bind to the newly forming RNA. In an ATP-dependent process, rho travels along the RNA at a speed comparable to the transcription process itself (fig. 10.11). Possibly, when RNA polymerase pauses at the stem-loop structure, rho catches up to the polymerase and unwinds the DNA-RNA hybrid, leting the DNA, RNA, and polymerase fall free. Rho can do this because it has DNA-RNA helicase (unwinding) properties.

The process of transcription termination is probably more complex than described. Significant interactions may take place with other proteins, and particular sequences surrounding the termination sequence may also be significant in the termination process. This is an area of active research.

Figure 10.12 shows an overview of transcription. The information of a gene, coded in the sequence of nucleotides in the DNA, has been transcribed into a complementary sequence of nucleotides in the RNA. This RNA transcript contains a complement of the template strand of the gene's DNA and thus acts as a messenger from the gene to the cell's protein-synthesizing complex. The transcript contains nucleotide sequences that will be translated into amino acids—coding segments—as well as noncoding segments before and after. The translatable segment, or gene, almost always begins with a three-base sequence, AUG, which is known as an initiator codon, and ends with one of the three-base sequences, UAA, UAG, or UGA, known as nonsense codons. (We discuss these signals in chapter 11.)

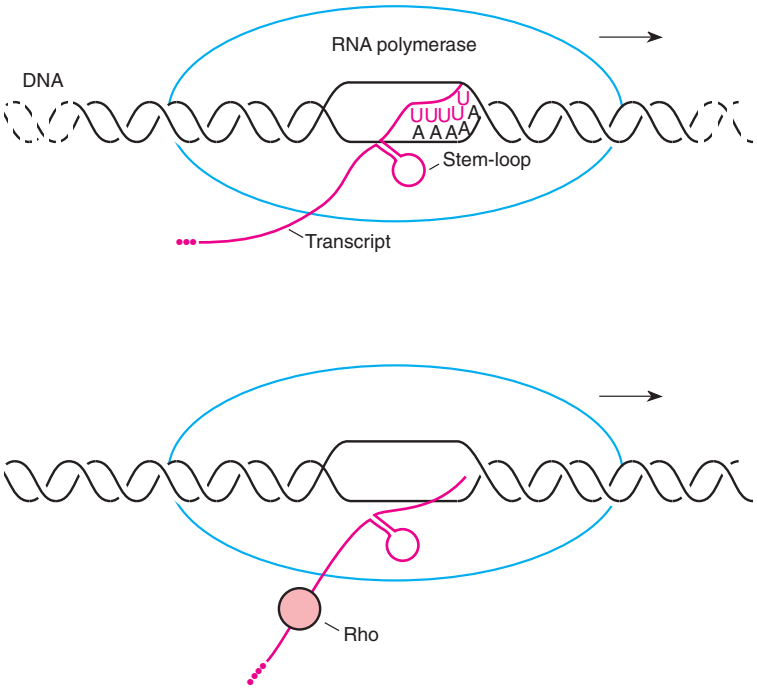


Figure 10.11 Rho-independent (*top*) and rho-dependent (*bottom*) termination of transcription are preceded by a pause of the RNA polymerase at a terminator sequence. Presumably, the stem-loop structure in the nascent RNA causes the pause in both cases.

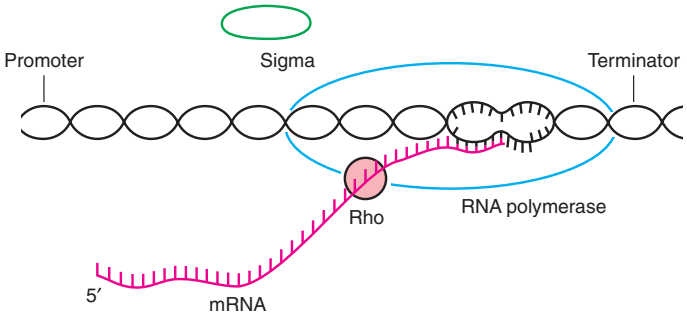


Figure 10.12 Transcription overview and RNA polymerase molecules. RNA polymerase is transcribing near the terminator. The rho factor—actually made up of six subunits—is shown on the newly formed RNA. The sigma factor is shown nearby, detached from the core polymerase.

The portion of the RNA transcript that begins at the start of transcription and goes to the translation initiator codon (AUG) is referred to as a **leader,** or 5′ untranslated sequence. The length of RNA from the nonsense codon (UAA, UAG, or UGA) to the last nucleotide transcribed is the **trailer,** or 3′ untranslated sequence. These sequences play a role in recognizing messenger RNA and ensuring its structural stability at the ribosome during the process of translation; the leader region can also have regulatory functions (see chapter 14). Figure 10.13 diagrams a complete prokaryotic RNA transcript. In this simplified drawing, the transcript has only one gene (AUG → UAA). However, the average prokaryotic transcript contains the information for several genes. We will say more about the parts of a transcript later in this chapter and the next. Now we turn our attention to the types of transcripts: ribosomal, transfer, and messenger RNA.

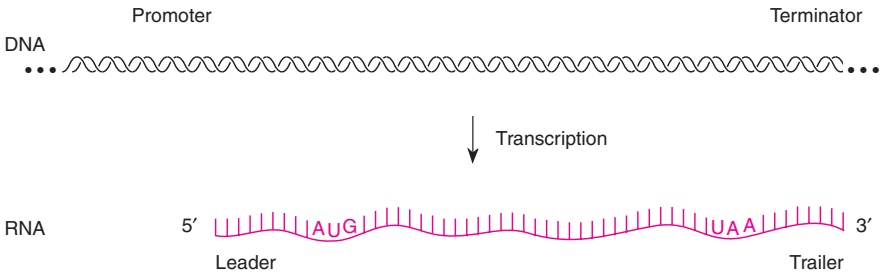


Figure 10.13 Transcribed piece of prokaryotic RNA and its DNA template region. Note the promoter and terminator regions on the DNA and the leader and trailer regions on the RNA. The initiation (AUG) and nonsense (UAA) codons for protein synthesis are shown. These signals are read at the ribosome at the time of translation.

RIBOSOMES AND RIBOSOMAL RNA

Ribosomes are organelles in the cell, composed of proteins and RNA (ribosomal RNA, or rRNA), where protein synthesis occurs. In a rapidly growing *E. coli* cell, ribosomes can make up as much as 25% of the mass of the cell. Ribosomes, as well as other small particles and molecules, are measured in units that describe their rate of sedimentation during density-gradient centrifugation in sucrose. This technique gives information on size and shape (due to the speed of sedimentation) while simultaneously isolating the molecules. Isolation by centrifugation in sucrose is a relatively gentle isolation technique; the molecules still retain their biological properties and can be used for further experimentation. In the 1920s, physical chemist T. Svedberg developed ultracentrifugation, giving his name to the unit of sedimentation: the **Svedberg unit,** S.

In sucrose density-gradient centrifugation, the gradient is formed by layering on decreasingly concentrated sucrose solutions. In a related technique, cesium chloride density-gradient centrifugation, mentioned in chapter 9, the gradient develops during centrifugation. The sucrose centrifugation is stopped after a fixed time, whereas in the cesium chloride technique, the system spins until it reaches equilibrium. The sucrose method tends to be more rapid. Samples can be isolated from a sucrose gradient by punching a hole in the bottom of the tube and collecting the drops in sequentially numbered containers. The first (lowest-numbered) containers will contain the heaviest molecules (with the highest S values).

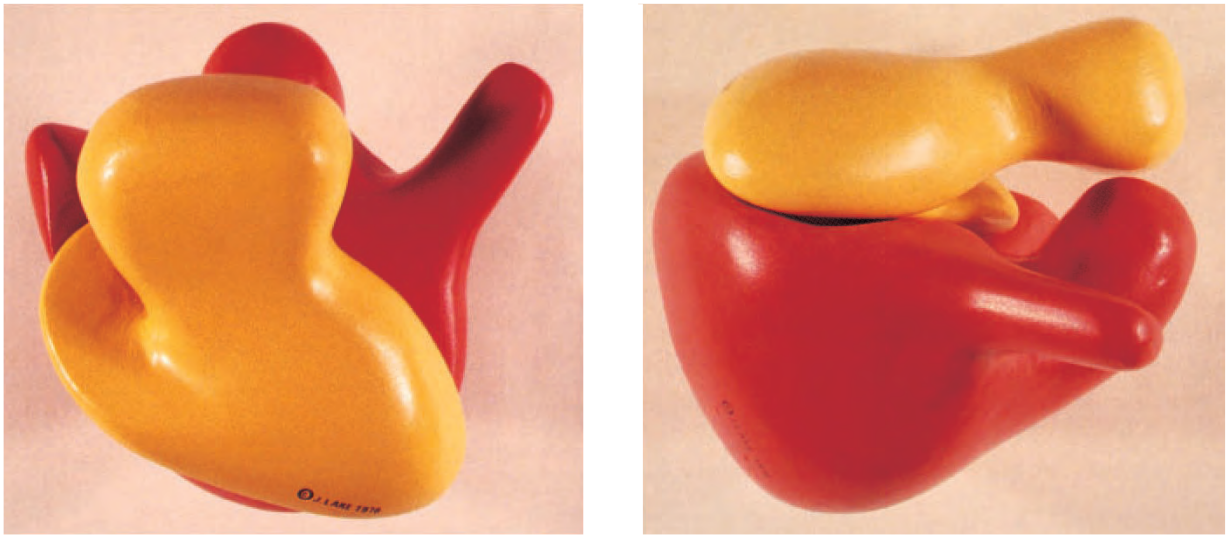
Ribosomes in all organisms are made of two subunits of unequal size. The sedimentation value is 50S (Svedberg units) for the large one in *E. coli* and 30S for the smaller one. Together they sediment at about 70S. Eukaryotic ribosomes vary from 55S to 66S in animals and 70S to 80S in fungi and higher plants. Most of our discussion will be confined to the well-studied ribosomes of *E. coli.*

Each ribosomal subunit comprises one or two pieces of ribosomal RNA and a fixed number of proteins. The 30S subunit of *E. coli* has twenty-one proteins and a 16S molecule of ribosomal RNA, and the 50S subunit has thirty-four proteins and two pieces of ribosomal RNA—one 23S and one 5S section (fig. 10.14). Advances in understanding ribosomal structure have come about after protein chemists isolated and purified all the proteins of the ribosome. This allowed researchers to experiment on the proper sequence needed to assemble the subunits and also allowed them to develop immunological techniques to show the positions of many proteins in the completed ribosomal subunits.

In *E. coli,* all three ribosomal RNA segments are transcribed as a single long piece of RNA that is then cleaved and modified to form the final three pieces of RNA (16S, 23S, and 5S). The region of DNA that contains the three ribosomal RNA molecules also contains genes for four transfer RNAs (fig. 10.15). There appear to be about five to ten copies of this region in each chromosome of *E. coli.* The occurrence of the three ribosomal RNA segments on the same piece of RNA ensures a final ratio of 1:1:1, the ratio needed for ribosomal construction.

TRANSFER RNA

During protein synthesis (see fig. 10.3), a messenger RNA, carrying the information transcribed from the gene (DNA), is bound to the ribosome. Amino acids are brought to the ribosome attached to transfer RNAs. The code is read in sequences of three nucleotides, called **codons.** The nucleotides of the codon on messenger RNA are complementary to and pair with a sequence of three bases—the **anticodon**—on a transfer RNA. Each different transfer RNA carries a specific amino acid. Thus, the transfer RNA recognizes the specificity of the genetic code (fig. 10.16).



(a) **30S**
Twenty-one proteins
One 16S rRNA

(b) **50S**
Thirty-four proteins
One 23S rRNA
One 5S rRNA

Figure 10.14 The *E. coli* ribosome. (*a*) and (*b*) show models of the 70S ribosome of *E. coli,* revealing the relationship of the small (*yellow*) and large (*red*) subunits at the time of translation. The 30S ribosomal subunit is composed of twenty-one proteins and one 16S piece of ribosomal RNA. The 50S subunit is composed of thirty-four proteins and two pieces of ribosomal RNA, 23S and 5S. ([*a* and *b*] James A. Lake, *Journal of Molecular Biology* 105 (1976):131–59. Reproduced by permission of Academic Press.)

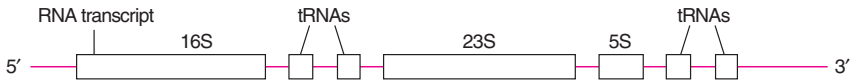


Figure 10.15 The *E. coli* transcript that contains the three ribosomal RNA segments also contains four tRNAs and some spacer RNA (*red*) that separates the tRNA and rRNA genes.

The correct amino acid is attached to its transfer RNA by one of a group of enzymes called **aminoacyl-tRNA synthetases.** One specific aminoacyl synthetase exists for every amino acid, but the synthetase may recognize more than one transfer RNA because there are more transfer RNAs (and codons) than there are amino acids. (In chapter 11 we discuss the genetic code in more detail.) R. W. Holley, a Nobel laureate, and his colleagues were the first to discover the nucleotide sequence of a transfer RNA; in 1964, they published the structure of the alanine transfer RNA in yeast (fig. 10.17). The average transfer RNA is about eighty nucleotides long.

Similarities of All Transfer RNAs

Transfer RNAs have several unusual properties. For one, all the different transfer RNAs of a cell have the same general shape; when purified, the heterogeneous mixture of all of a cell's transfer RNAs can form very regular crystals. The regularity of the shape of transfer RNAs makes sense. During the process of protein synthesis, two transfer RNAs attach next to each other on a ribosome, and a peptide bond forms between their amino acids. Thus, any two transfer RNAs must have the same general dimensions as well as similar structures so that they can be recognized and positioned correctly at the ribosome.

An obvious feature of the transfer RNA in figure 10.17 is that it has **unusual bases.** When this transfer RNA is originally transcribed from DNA, it is about 50% longer than the final eighty nucleotides. In fact, some transcripts contain two copies of the same transfer RNA, or sometimes several different transfer RNA genes are part of the same transcript (see fig. 10.15). The original transcription of transfer RNAs is completely regular: It does not involve unusual bases. The transcript is then processed down to the final size of a transfer RNA by various nucleases that remove trailing and leading pieces of RNA. In eukaryotes, a CCA sequence of nucleotides is added at the 3′ end by a nucleotidyl

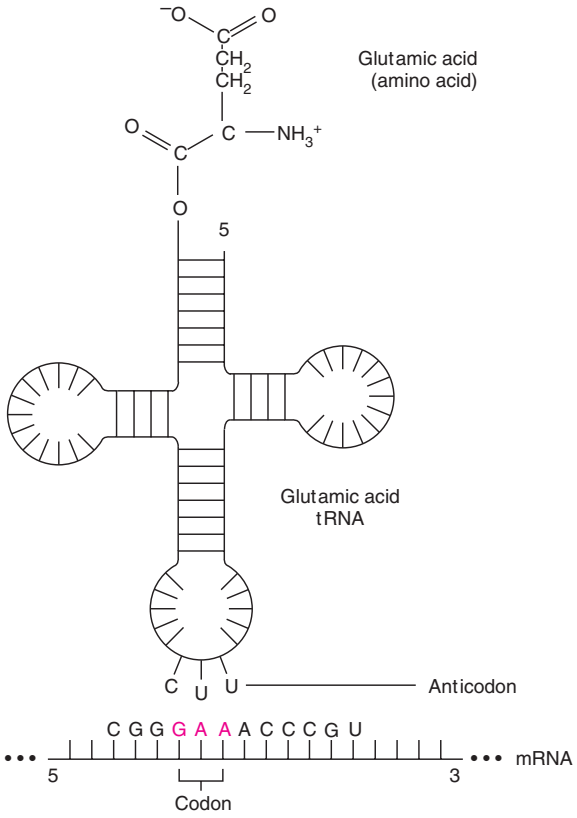


Figure 10.16 Specificity of the genetic code manifests itself in the transfer RNA, in which a particular anticodon is associated with a particular amino acid. In this case, glutamic acid is attached to its proper transfer RNA, which has the anticodon CUU.

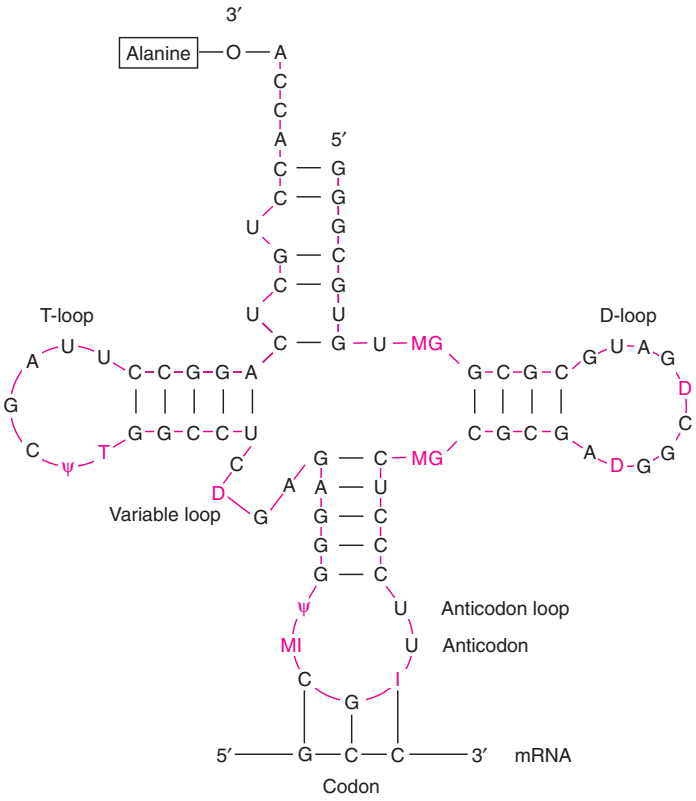


Figure 10.17 Structure and sequences of alanine transfer RNA in yeast. Note the modified bases in the loops. The anticodon of the transfer RNA is shown paired with its complementary codon in the DNA. (Source: Data from R. W. Holley, et al., "Structure of a ribonucleic acid," *Science*, 147:1462–65, 1965.)

transferase enzyme. Then the transfer RNA is further modified, frequently by the addition of methyl groups to the bases already in the RNA (fig. 10.18). Presumably, these unusual bases disrupt normal base pairing and are in part responsible for the loops the unpaired bases form (see fig. 10.17).

Transfer RNA Loops

It is believed that the first loop on the 3′ side (the T- or T-ψ-C-loop) is involved in making the transfer RNA recognizable to the ribosome. The ribosome must hold each transfer RNA in the proper orientation to check the complementarity of the anticodon of the transfer RNA and the codon of the messenger RNA. The center loop of transfer RNA is the anticodon loop. The aminoacyl-tRNA synthetases seem to recognize many points all over the transfer RNA molecule (see chapter 11).

The amino acid is attached to the ACC sequence on the 3′ end of the transfer RNA. The ribosome-binding loop on all transfer RNAs has the T-ψ-C-G sequence. The anticodon on all is bounded by uracil on the 5′ side and a purine on the 3′ side. Thus, there is a good deal of general similarity among all the transfer RNAs, consistent with the fact that they all enter protein synthesis in the same way. The actual shape of the functional transfer RNA in the cell is not an open cloverleaf, as shown in figure 10.17; rather, the whole molecule exhibits helical twisting due to pairing of complementary regions (fig. 10.19).

Earlier we considered a rough definition of a gene as a length of DNA that codes for one protein. But we have just encountered an inconsistency—genes code for both transfer RNAs and ribosomal RNAs, yet neither is eventu-

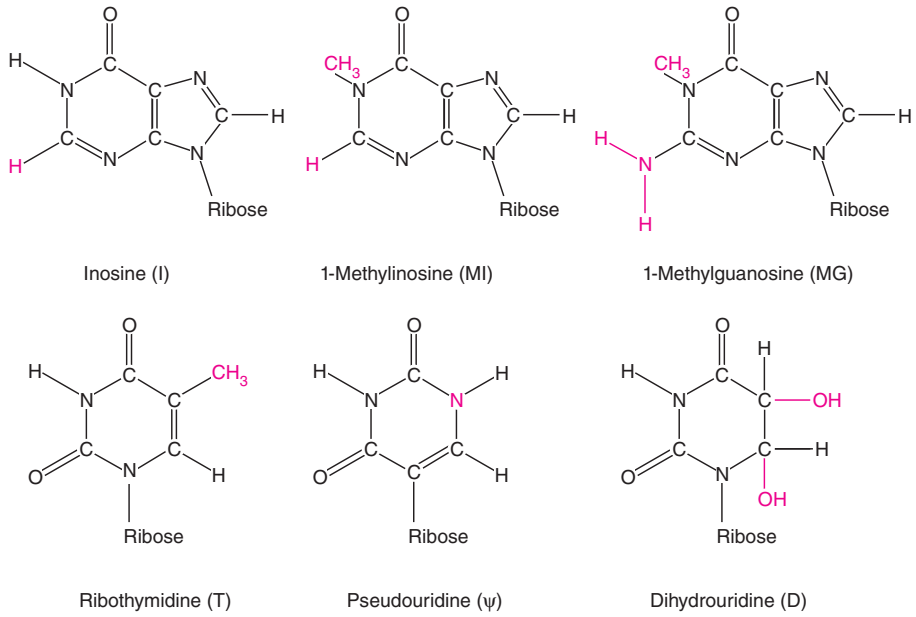


Figure 10.18 Structures of the modified bases found in alanine transfer RNA of yeast. The various modifications of normal bases are shown in *red*.

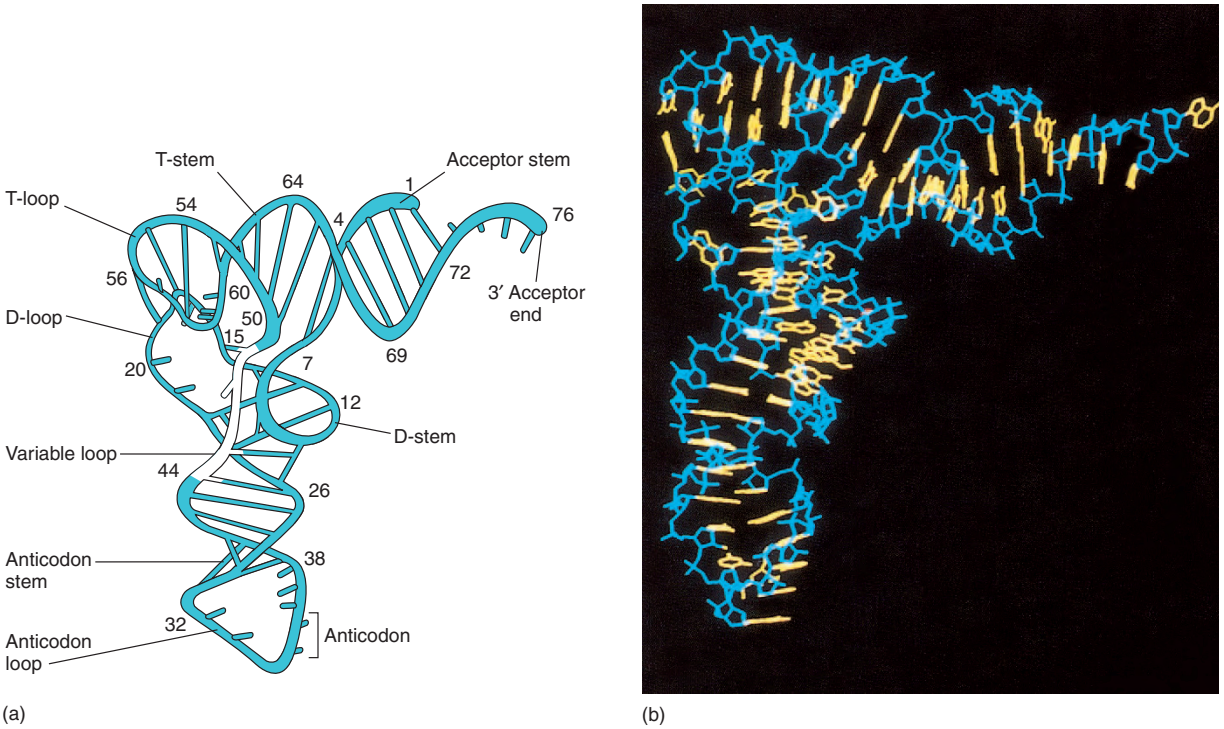


Figure 10.19 Structure of yeast phenylalanine transfer RNA. (*a*) A diagram showing coiling of the sugar-phosphate backbone. (*b*) A molecular model with bases in *yellow* and backbone in *blue*. The two parts of the figure (*a* and *b*) are in the same orientation.
([*b*] Courtesy of Alexander Rich.)

ally translated into a protein. Their transcripts function as final products without ever being translated. Thus, transfer RNA and ribosomal RNA are the major exceptions to the general rule that a gene codes for a protein.

EUKARYOTIC DNA TRANSCRIPTION

The Nucleolus in Eukaryotes

Eukaryotes have four segments of ribosomal RNA in the ribosome, compared with three in prokaryotes. The smaller ribosomal subunit has an 18S piece of RNA, and the larger subunit has 5S, 5.8S, and 28S segments. All but the 5S ribosomal RNA section are transcribed as part of the same piece of RNA. However, eukaryotic cells have many copies of these ribosomal RNA genes, depending on the species. For example, the fruit fly, *Drosophila melanogaster,* has about 130 copies of the DNA region that the larger segments of ribosomal RNA are transcribed from. These regions occur in tandem on the sex (X and Y) chromosomes and are known collectively as the nucleolar organizer (see chapter 3). The smallest ribosomal RNA subunit is also produced from a duplicated gene, but at a different point in the genome. For example, in *D. melanogaster,* the 5S subunit is produced on chromosome 2.

Eukaryotes—unlike prokaryotes, which have only one RNA polymerase—have three RNA polymerases. Eukaryotic RNA polymerase I (or polymerase A) transcribes only the nucleolar organizer DNA. RNA polymerase II (or polymerase B) transcribes most genes. RNA polymerase III (or polymerase C) transcribes small genes, primarily the 5S ribosomal RNA gene and transfer RNA genes (table 10.3). In addition, mitochondria, chloroplasts, and some phages have other RNA polymerases.

Table 10.3 **Prokaryotic and Eukaryotic RNA Polymerases**

Enzyme	Function
Prokaryotic	
RNA polymerase	Transcribes DNA template
Primase	Primer synthesis during DNA replication
Eukaryotic	
RNA polymerase I	Transcribes nucleolar organizer
RNA polymerase II	Transcribes most genes
RNA polymerase III	Transcribes 5S rRNA and tRNA genes
Primase	Primer synthesis during DNA replication

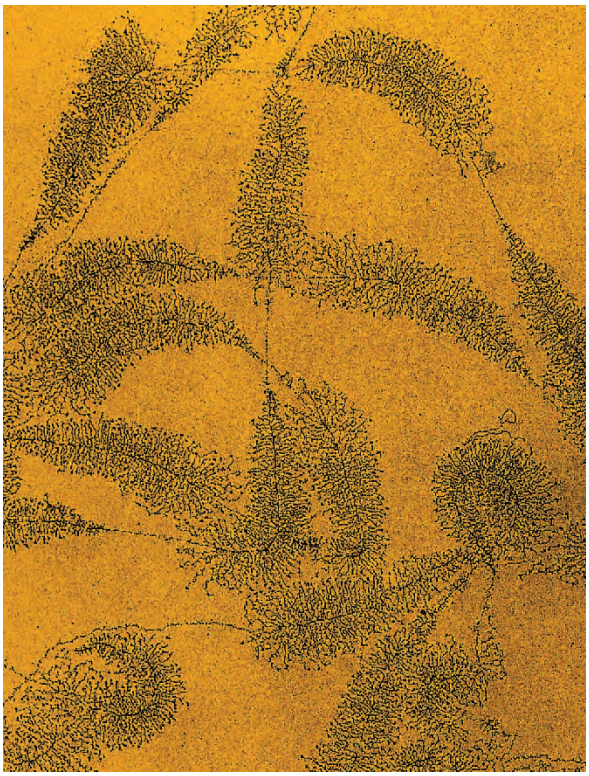


Figure 10.20 Transcription in the nucleolus of the newt, *Triturus*. Tandem repeats of the large ribosomal RNA genes are being transcribed. The polarity of the process (progressing from small to large transcripts), as well as the spacer DNA (*thin lines* between transcribing areas), is clearly visible. Magnification 18,000×. (© O. L. Miller, B. R. Beatty, D. W. Fawcett/Visuals Unlimited.)

At the nucleolar organizer, the nucleolus forms the familiar dark blob found in eukaryotic nuclei. The nucleolus is the place where ribosomes are assembled. The various ribosomal proteins that have been manufactured in the cytoplasm migrate to the nucleus and eventually to the nucleolus, where, with the final forms of the ribosomal RNAs, they are assembled into ribosomes.

In the nucleolar organizer, an untranscribed region of spacer DNA separates each repeat of the large ribosomal RNA gene. This is shown in figure 10.20 and diagrammed in figure 10.21. In the electron micrograph in figure 10.20, the polarity of transcription is evident from the short RNA at one end of the transcribing segment and the long RNA at the other end, with a uniform gradation between. Notice that many RNA polymerases are transcribing each region at the same time. The regions between the transcribed DNA segments are the spacer DNA regions.

Like transfer RNAs, ribosomal RNAs are also modified: some uridines are converted to pseudouridines, and some

ribose sugars are methylated. These conversions take place in the nucleolus, orchestrated by particles composed of small RNA segments and protein. The RNA segments are referred to as **small nucleolar RNAs (snoRNAs)** and, when combined with protein, are referred to as **small nucleolar ribonucleoprotein particles (snoRNPs).** Each different snoRNP has a snoRNA that is complementary to the regions surrounding the nucleotide to be modified. Thus, sites for modification are chosen based on complementarity to a snoRNA, which then somehow directs the modification to take place.

Differences Between Eukaryotic and Prokaryotic Transcription

Although all aspects of transcription differ to some extent between prokaryotes and eukaryotes, we will look at two major differences here: the coupling of transcription and translation that is possible in prokaryotes, and the extensive **posttranscriptional modifications** that occur in eukaryotic messenger RNA. In *E. coli,* translation of the newly transcribed messenger RNA into a protein can take place before transcription is complete (fig. 10.22). The messenger RNA is synthesized in the 5′ → 3′ direction, and it is

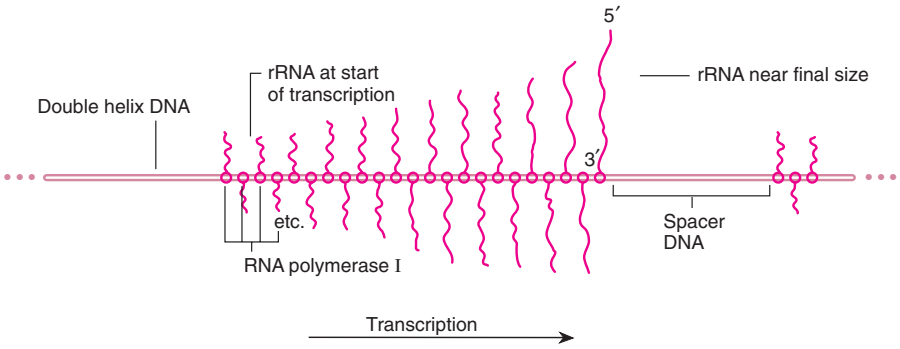
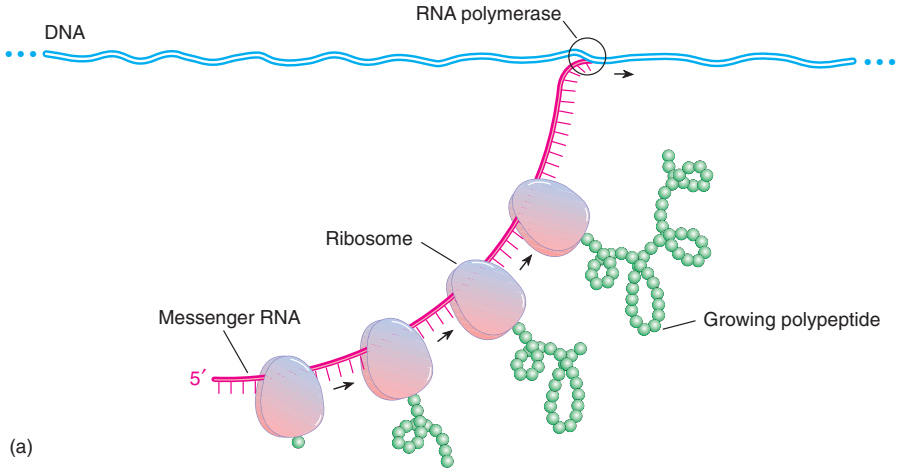
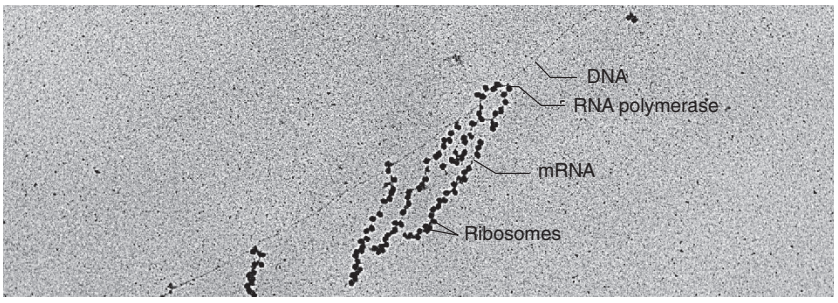


Figure 10.21 Details of the transcription of the large ribosomal RNA genes shown in figure 10.20. Note the polarity of the process and the spacer DNA, as seen in figure 10.20.



(a)



(b)

Figure 10.22 (*a*) In prokaryotes, translation of messenger RNA by ribosomes begins before transcription is complete. Ribosomes attach to the growing mRNA strand when the 5′ end becomes accessible. They then move along the RNA as it elongates. When the first ribosome moves from the 5′ end, a second ribosome can attach, and so on. (*b*) Electron micrograph of events diagrammed in (*a*). The growing polypeptides cannot be seen in this preparation. Magnification 44,000×. ([*b*] Courtesy of O. L. Miller, Jr.)

near the 5′ end that translation begins. As soon as the 5′ end of the RNA is available, a ribosome can attach to the messenger RNA and move along it in the 5′ → 3′ direction, lengthening the growing polypeptide as it moves. When the first ribosome moves away from the 5′ end of the transcript, a second ribosome can attach and begin translation. These processes are repetitive, as electron micrographs (fig. 10.22*b*) clearly show. In eukaryotes, however, messenger RNA is synthesized in the nucleus, but protein synthesis takes place in the cytoplasm. (This regional division of labor is not present in *E. coli* because, among other reasons, the bacterium has no nucleus.) Before a eukaryotic messenger RNA leaves the nucleus, it is highly modified by processes that generally do not occur in prokaryotes.

Promoters

Eukaryotic promoters are somewhat similar to prokaryotic promoters; both are regions of DNA at the beginnings of genes with signals that allow RNA polymerase to attach and begin transcription. In eukaryotes, however, more proteins are involved in promoter recognition, and more proteins are involved in the control of transcription, many recognizing signals thousands of base pairs away. We discuss these control processes in eukaryotes in chapter 16.

All three eukaryotic RNA polymerases (I, II, and III) recognize a seven-base sequence, TATAAAA, located at about −25 on the promoter DNA. It is similar to the −10 sequence in prokaryotes and is called the **TATA box** (or **Hogness box** after its discoverer, D. Hogness). Since RNA polymerase II transcribes most of the genes in eukaryotes, we turn our attention specifically to it.

Among the large number of promoters that have been sequenced, a few lack the TATA box, yet are still transcribed. Transcription initiation in these promoters appears to be controlled by a CT-rich area, called the **initiator element (Inr),** at +1 of the transcript (close to the transcription start site), coupled with a **downstream promoter element (DPE)** at about +28 to +34 of the transcript. In TATA-less promoters, a protein called TFIID requires both these elements to bind. The initiator element has a consensus sequence of TCA(G or T)T(T or C), and the downstream promoter element has the consensus sequence of (A or G)G(A or T)CGTG. We will concentrate on RNA polymerase II genes with TATA boxes.

Yeast RNA polymerase II is a protein of twelve subunits. This enzyme cannot locate promoters or attach to DNA in a stable fashion. To attach at the beginnings of genes, RNA polymerase II must interact with several proteins called **general transcription factors.** In eukaryotes, general transcription factors are named after the polymerase they work with. Thus, the transcription factor that recognizes the TATA box for polymerase II genes is called TFIID (D being the fourth letter of the alphabet for the fourth transcription factor so named). TFIID is composed of one

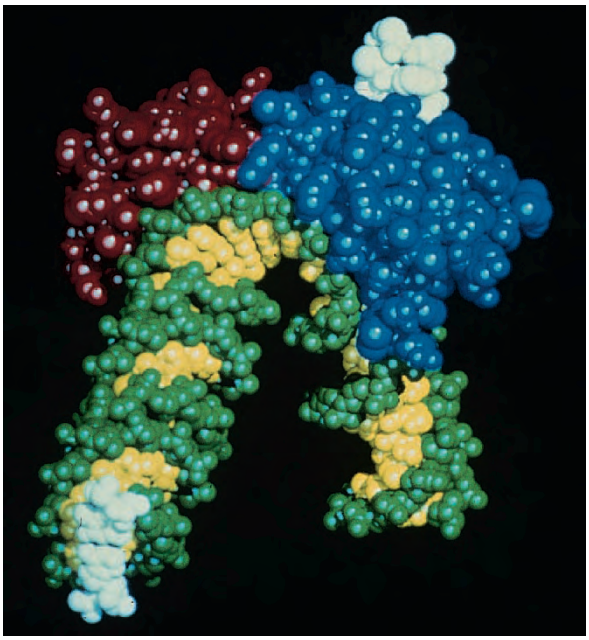


Figure 10.23 Molecular space-filling model of a yeast TATA-binding protein attached to a TATA box on the DNA. The DNA sugar-phosphate backbone is *green* and the bases are *yellow.* The protein has twofold symmetry (*red* and *blue*). Note the bending of the DNA through 80 degrees, which also opens up the minor groove of the DNA. The *upper white* atoms are the N-terminus of the TATA-binding protein; the *lower white* atoms are the first base pair at which transcription begins. (Courtesy of J. L. Kim and S. K. Burley. From J. L. Kim, J. H. Geiger, S. Hahn, and P. B. Sigler, "Crystal Structure of a Yeast TBP TATA-box Complex." *Nature* 365 (6446): 520–27, Oct. 7, 1993. © Macmillan Magazines, Ltd. Figure adapted from the work of S. K. Burley.)

subunit that recognizes the TATA sequence, called **TATA-binding protein (TBP),** and up to a dozen other proteins called **TBP-associated factors (TAFs),** which recognize the initiator element, when present, and aid in regulating transcription. TFIID is, in essence, similar to the sigma factors of prokaryotic RNA polymerase. One interesting aspect of the binding of TBP is that it causes a significant bending and opening of the DNA (fig. 10.23). This bending may be an important signal for other binding proteins.

Once TFIID binds to the TATA box, a cascade of recruitment (binding) of other transcription factors takes place. Transcription factors IIA, IIB, and IIF bind, as does RNA polymerase II in an unphosphorylated state. Then transcription factors IIE and IIH bind, forming a **preinitiation complex (PIC),** equivalent to the *E. coli* holoenzyme (fig. 10.24*a*). The RNA polymerase II is then phosphorylated, presumably by TFIIH, which is a kinase; at this point, most of the transcription factors drop off, leaving the **elongation complex,** which carries out a

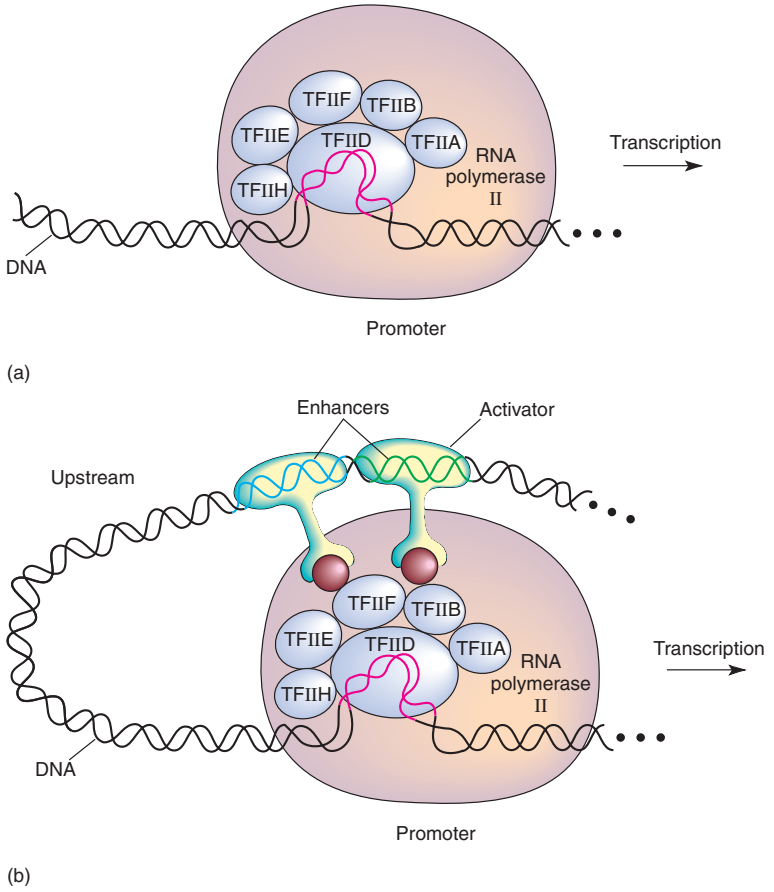


Figure 10.24 (*a*) An RNA polymerase II preinitiation complex at a promoter. TFIID binds to the TATA box (*red*). The other transcription factors are then recruited with the polymerase. (*b*) Two activators (*yellow*) are shown bound at one end (their DNA domains) to enhancers (*blue* and *green*) upstream on the DNA. The activators are bound at their other ends (their transcriptional activation domains) to other proteins associated with the polymerase machinery. Phosphorylation of the polymerase initiates activated transcription.

Table 10.4 **Putative Roles of the General Transcription Factors of RNA Polymerase II**

General Transcription Factor	Function
TFIID, TBP	Recognizes TATA box
TFIID, TAFs	Recognizes initiator element and regulatory proteins
TFIIA	Stabilizes TFIID
TFIIB	Aids in start-site selection by RNA polymerase II
TFIIE	Controls TFIIH functions; enhances promoter melting
TFIIF	Destabilizes nonspecific interactions of RNA polymerase II and DNA
TFIIH	Melts promoter with helicase activity; activates RNA polymerase II with kinase activity

Source: Data from R. G. Roeder, "The Role of General Initiation Factors in Transcription by RNA Polymerase II" in *Trends in Biochemical Sciences,* 21:327–35, 1996.

basal rate of transcription (fig. 10.25). TFIIH also has a role here, since it is also a helicase. Table 10.4 summarizes the postulated roles of the general transcription factors.

For activated transcription, a high level of transcription, to take place, other factors are needed that are involved in controlling which promoters are actively transcribed. These other factors are **activators** or **specific transcription factors** that bind to DNA sequences called **enhancers.** Enhancers are often hundreds or thousands of base pairs upstream from the promoter (fig. 10.24*b*).

Note that much of this information has been gathered by footprinting, mutational studies, cloning and isolating the genes and proteins involved, and then reconstituting various purified combinations in the test tube. These studies are combined with kinetic research to determine which arrangements are stable, immunological research

to isolate various components with antibodies, and photocrosslinking studies to determine which moities are in contact with each other.

These specific transcriptional activators have domains (regions) that recognize their specific enhancer sequences, regions that recognize proteins associated with the polymerase (general transcription factors), and regions that allow the joint attachment of other transcription factors (fig. 10.24*b*). Similar to activators and enhancers, repressors can bind to silencer regions of DNA, often far upstream of the promoters, to repress transcription. Thus, many genes are associated with numerous and complex arrangements of transcription factors, providing elaborate control of transcription (see chapter 16).

For specific transcription factors to attach to both enhancers and the polymerase machinery, possibly thousands of base pairs apart, the DNA must bend to allow them to come into the range of the polymerase. Electron micrographs clearly show this DNA bending and looping (fig. 10.26).

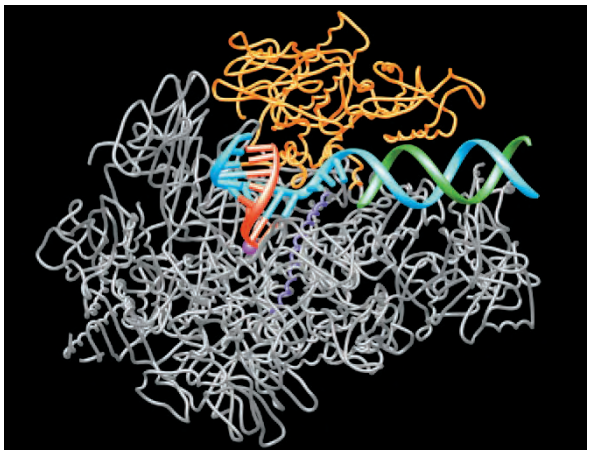
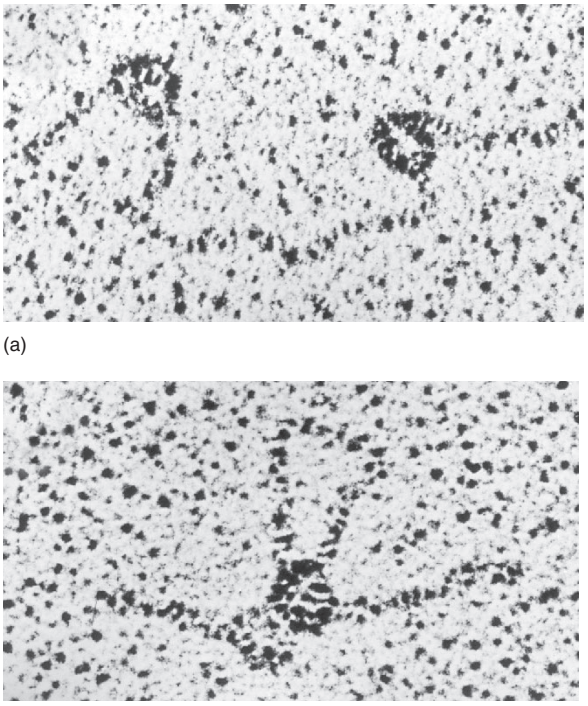


Figure 10.25 The RNA polymerase II elongation complex with part of the protein structure removed to show the DNA and RNA within the cleft of the protein. The DNA is *blue* (template strand) and *green* (nontemplate strand) with the RNA *red*. The majority of protein is shown as *gray*; the part in *yellow* is a domain that appears to open for DNA loading and is in a closed state during elongation, thus acting as a clamp on the DNA and RNA. Closure of the clamp allows for the high stability of transcribing complexes and thus for processivity of the polymerase. The *purple* part is a helix that crosses the major cleft of the enzyme. The DNA template strand is led over this helix towards the active site. The *pink* sphere is a magnesium ion in the active site, where RNA synthesis occurs. (P. Cramer, D. A. Bushnell and R. D. Kornberg. RNA polymerase II at 2.8A resolution and A. L. Gnatt, P. Cramer, J. Fu, D. A. Bushnell and R. D. Kornberg, Structure of an RNA polymerase II transcribing complex. Reprinted by permission of the authors.)

Although RNA polymerases I and III seem to have termination signals similar to rho-independent promoters in prokaryotes, termination of transcription of RNA polymerase II genes is more complex, coupled with further processing of the mRNA.

Before we move on, several other points merit discussion. First, unlike prokaryotic RNA polymerases, eukaryotic RNA polymerases do proofread (showing $3' \rightarrow 5'$ exonuclease activity). Second, as we will discuss in chapter 15, eukaryotic DNA is complexed with histone proteins that can interfere with transcription. In turn, part of the RNA polymerase II complex is made up of proteins that can disrupt the histones bound to the DNA.

In addition, the RNA polymerase II complex contains proteins that act as mediators between activators and the polymerase holoenzyme. This complex coordination of the initiation of transcription in eukaryotes has been termed **combinatorial control;** the huge initiation complex may contain 85 or more different polypeptides.



(a)

(b)

Figure 10.26 The interaction between an activator and RNA polymerase (in this case, in prokaryotes). (*a*) In this system, the RNA polymerase of *E. coli* (the more heavily stained sphere) is controlled by an activator called NtrC (the more lightly stained sphere). The activator is bound to an enhancer, and the polymerase is bound to the promoter. (*b*) The activator has bound to the polymerase, causing a looping of the DNA. Compare with figure 10.24. (Courtesy of Sydney Kustu.)

Finally, transcription in the archaea, although under much simpler control than in the eukaryotes, resembles transcription in eukaryotes rather than prokaryotes.

The study of the details of the transcription process—its initiation, control, and termination—is one of the most active and exciting areas in modern genetics.

Caps and Tails

Eukaryotic transcription results in a **primary transcript.** In contrast to most prokaryotic transcripts that contain information from several genes, virtually all transcripts from higher eukaryotes contain the information from just one gene. (Transcripts from several genes are found in some lower eukaryotes, such as nematode worms.) Three major changes occur in primary transcripts of RNA polymerase II before transport into the cytoplasm: modifications to the 5′ and 3′ ends and removal of intervening sequences. We refer to these changes as posttranscriptional modifications.

At the 5′ end of polymerase II transcripts, 7-methyl guanosine is added in the "wrong" direction, 5′ → 5′ (fig. 10.27). This **cap** allows the ribosome to recognize the beginning of a messenger RNA. At the other end, the 3′ end of polymerase II transcripts, a sequence of twenty to two hundred adenine-containing nucleotides, known as a **poly-A tail,** is added by the enzyme poly-A polymerase. Polyadenylation takes place after the 3′ end of the transcript is removed by a nuclease that cuts about twenty nucleotides downstream from the signal 5′-AAUAAA-3′. The tail adds stability to the molecule and aids in its transportation from the nucleus.

When messenger RNAs were first studied in eukaryotes, the messenger RNAs in the nucleus were found to be much larger than those in the cytoplasm and were called **heterogeneous nuclear mRNAs,** or **hnRNAs.** It now turns out that these were primary transcripts, RNAs that had not had any of the major posttranscriptional modifications. In essence, they were premessenger RNAs.

Introns

Eukaryotes have segments of DNA within genes that are transcribed into RNA but never translated into protein sequences. These **intervening sequences,** or **introns,** are removed from the RNA in the nucleus before its transport into the cytoplasm (fig. 10.28). P. Sharp and his colleagues at MIT and R. Roberts, T. Broker, L. Chow, and their colleagues at the Cold Spring Harbor Laboratory first discovered introns in 1977. (Sharp and Roberts were awarded 1993 Nobel prizes for their work.) An example of a gene with introns appears in figure 10.29. The segments of the gene between introns, which are transcribed and translated—and hence exported to the cytoplasm and expressed—are termed **exons.** The results of intron removal are clear when a messenger RNA with its introns removed is hybridized with the original gene (fig. 10.30). The DNA forms double-stranded structures with the exons in RNA. The introns in DNA have nothing to pair with in the RNA, so they form single-stranded loops. Introns also occur in eukaryotic transfer RNA and ribosomal RNA genes.

For introns to be removed, the ends of the exons must be brought together and connected in a process called *splicing.* At least two types of splicing occur, although they are related: self-splicing and protein-mediated splicing.

Self-Splicing

In 1982, Thomas Cech and his colleagues, building on the work of others, including Sidney Altman, who showed that RNA can have catalytic properties, discovered self-splicing by RNA. (Cech and Altman were awarded 1989

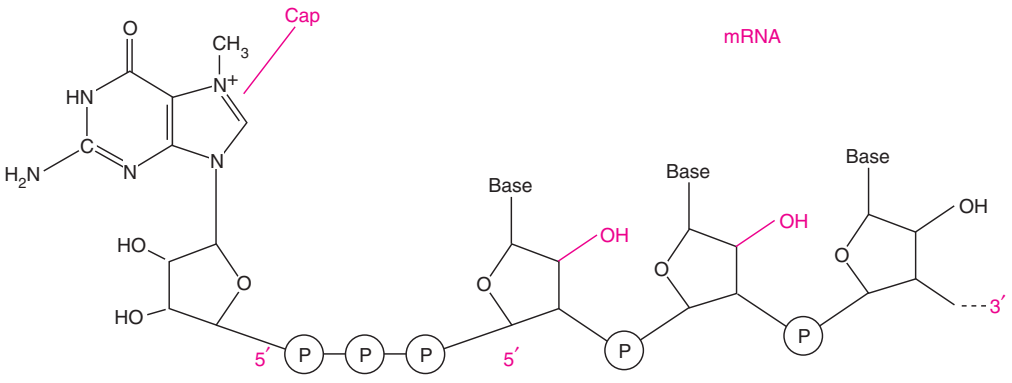


Figure 10.27 A cap of 7-methyl guanosine is added in the "wrong" direction (5′ → 5′), to the 5′ end of eukaryotic mRNAs. In some cases, the 2′ -OH groups on the second or second and third riboses (*red*) are methylated.

Richard J. Roberts (1943–).
(Courtesy of Richard J. Roberts.)

Philip A. Sharp (1944–).
(Courtesy of Dr. Philip A. Sharp.)

Thomas Broker (1944–).
(Courtesy of Dr. Thomas Broker.)

Louise T. Chow (1943–).
(Courtesy of Dr. Louise Chow.)

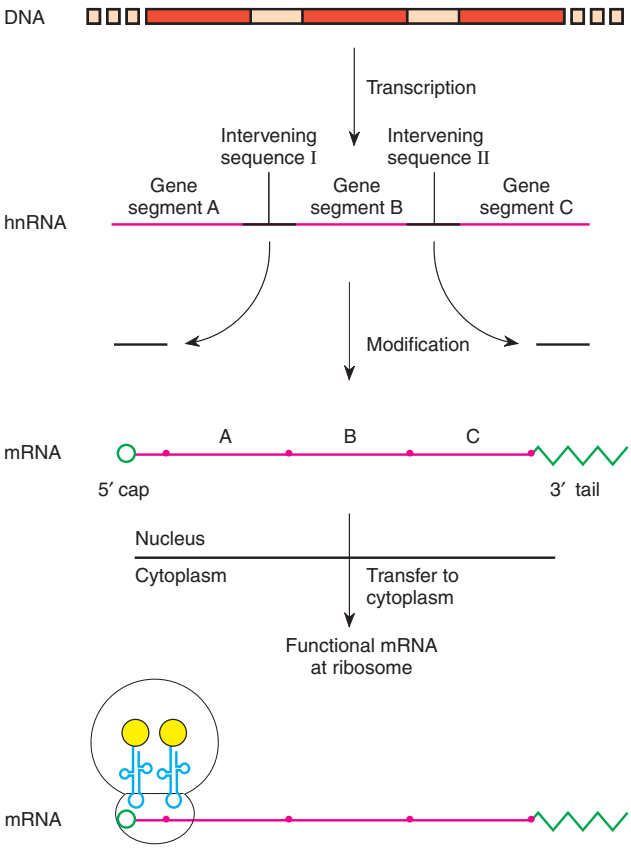


Figure 10.28 In eukaryotic DNA, intervening sequences, or introns, are removed from the RNA in the nucleus before the mRNA is transported into the cytoplasm and translated. Other modifications consist of splicing, 5′ capping, and 3′ polyadenylation.

Nobel prizes in chemistry.) Working with an intron in the 35S ribosomal RNA precursor in the ciliated protozoan, *Tetrahymena,* Cech and his colleagues found that they could induce intron removal in vitro with no proteins present. A guanine-containing nucleotide (GMP, GDP, or GTP) had to be present. Figure 10.31 diagrams how self-splicing occurs. The intron is acting as an enzyme; we call an RNA with enzymatic properties a **ribozyme.**

During self-splicing, the U-A bond at the left (5′) side of the intron is transferred to the GTP. The U that is now unbonded displaces the G at the right (3′) side of the intron, reconnecting the RNA with a U-U connection and releasing the intron (fig. 10.31). Since all bonds are reversible transfers (transesterifications) rather than new bonds, no external energy source is required. Self-splicing introns of this type are called **group I introns.** An extensive secondary structure (RNA stem-loops) that forms is also important in intron removal (box 10.3).

Although the first enzymatic activity of the ribozyme is its own removal, its secondary structure after removal gives it the ability to further catalyze reactions (fig. 10.32). The reactions that ribozymes catalyze are transesterifications and the hydrolysis reaction of splitting an RNA molecule into two parts. Ribozymes can also perform other functions, including peptide bond formation, covered in chapter 11. Currently, at least seven different classes of ribozymes are known, based on their enzymatic properties. A ribozyme that can split other RNAs and that occurs in small plant pathogens is called a **hammerhead ribozyme** (fig. 10.33) because of its shape. Because these RNA molecules are small, they have the potential to be modified in the laboratory for specific purposes related to clinical treatment and further study of RNA processing.

Self-splicing has also been found in genes in the mitochondria of yeast. These introns are referred to as **group II introns** because they use a different mechanism of splicing that does not require an external nucleotide. Instead, the first bond is transferred within the intron to an adenosine, forming a lariat structure (fig. 10.34). In order for the lariat to form, the ribose of the adenosine must make three phosphodiester bonds (fig. 10.35).

```
                  10        20        30        40        50        60        70       cAp      90
  0 GGCCAATCTGCTCACACAGGATAGAGAGGGCAGGAGCCAGGCAGAGCATATAAGGTGAGGTAGGATCAGTTGCTCCTCACATTTGCTTCTGACATAGTTG

100 TGTTGACTCACAACCCCAGAAACAGACATCATGGTGCACCTGACTGATGCTGAGAAGGCTGCTGTCTCTTGCCTGTGGGGAAAGGTGAACTCCATGAAG
                                                    MetValHisLeuThrAspAlaGluLysAlaAlaValSerCysLeuTrpGlyLysValAsnSerAspGluV

200 TTGGTGGTGAGGCCCTGGGCAGGTTGGTATCCAGGTTACAAGGCAGCTCAAGAAGAAGTTGGGTGCTTGGAGACAGAGGTCTGCTTTCCAGCAGACAC
    alGlyGlyGluAlaLeuGlyArg  30

300 TAACTTTCAGTGTCCCCTGTCTATGTTTCCCTTTTTAGGCTGCTGGTTGTCTACCCTTGGACCCAGCGGTACTTTGATAGCTTTGGAGACCTATCCTCTG
                                                  31  LeuLeuValValTyrProTrpThrGlnArgTyrPheAspSerLeuLysGlyTh

400 CCTCTGCTATCATGGGTAATGCCAAAGTGAAGGCCCATGGCAAGAAGGTGATAACTGCCTTTAACGATGGCCTGAATCACTTGGACAGCCTCAAGGGCAC
    laSerAlaIleMetGlyAsnAlaLysValLysAlaHisGlyLysLysValIleThrAlaPheAsnAspGlyLeuAsnHisLeuAspSerLeuLysGlyTh

500 CTTTGCCAGCCTCAGTGAGCTCCACTGTGACAAGCTGCATGTGGATCCTGAGAACTTCAGGGTGAGTCTGATGGGCACCTCCTGGGTTTCCTTCCCCTGC
    rPheAlaSerLeuSerGluLeuHisCysAspLysLeuHisValAspProGluAsnPheArg  104

600 TATTCTGCTCAACCTTCCTATCAGAAAAAAAGGGGAAGCGATTCTAGGGAGCAGTCTCCATGACTGTGTGTGGAGTGTTGACAAGAGTTCGGATATTTTA

700 TTCTCTACTCAGAATTGCTGCTCCCCCTCACTCTGTTCTGTGTTGTCATTTCCTCTTTCTTTGGTAAGCTTTTTAATTTCCAGTTGCATTTTACTAAATT

800 AATTAAGCTGGTTATTTACTTCCCATCCTGATATCAGCTTCCCCTCCTCCTTTCCTCCCAGTCCTTCTCTCTCTCCTCTCTCTTTCTCTAATCCTTTCCT

900 TTCCCTCAGTTCATTCTCTCTTGATCTACGTTTGTTTGTCTTTTTAAATATTGCCTTGTAACTTGCTCAGAGGACAAGGAAGATATGTCCCTGTTTCTTC

1000 TCATAGCTCAAGAATAGTAGCATAATTGGCTTTTATGCAGGGTGACAGGGGAAGAATATATTTTACATATAAATTCTGTTTGACATAGGATTCTTGTGGT

1100 GGTTTGTCCAGTTTAAGGTTGCAAACAAATGTCTTTGTAAATAAGCCTGCAGGTATCTGGTATTTTTGCTCTACAGTTATGTTGATGGTTCTTCCATATT

1200 CCCACAGCTCCTGGGCAATATGATCGTGATTGTGCTGGGCCACCTTGGCAAGGATTTCACCCCCGCTGCACAGGCTGCCTTCCAGAAGGTGGTGGCT
         105  LeuLeuGlyAsnMetIleValIleValLeuGlyHisHisLeuGlyLysAspPheThrProAlaAlaPheGlnLysValValAla

1300 GGAGTGGCCACTGCCTTGGCTCACAAGTACCACTAAACCCCCTTTCCTGCTCTTGCCTGTGAACAATGGTTAATTGTTCCCAAGAGAGCATCTGTCAGTT
     GlyValAlaThrAlaLeuAlaHisLysTyrHisTer
                                                                                    pA
1400 GTTGGCAAAATGATAGACATTTGAAAATCTGTCTTCTGACAAATAAAAAGCATTTATGTTCACTGCAATGATGTTTTAAATTATTTGTCTGTGTCATAGA

1500 AGGGTTTATGCTAAGTTTTCAAGATACAAAGAAGTGAGGGTTCAGGTCTCGACCTTGGGGAAATAAA
```

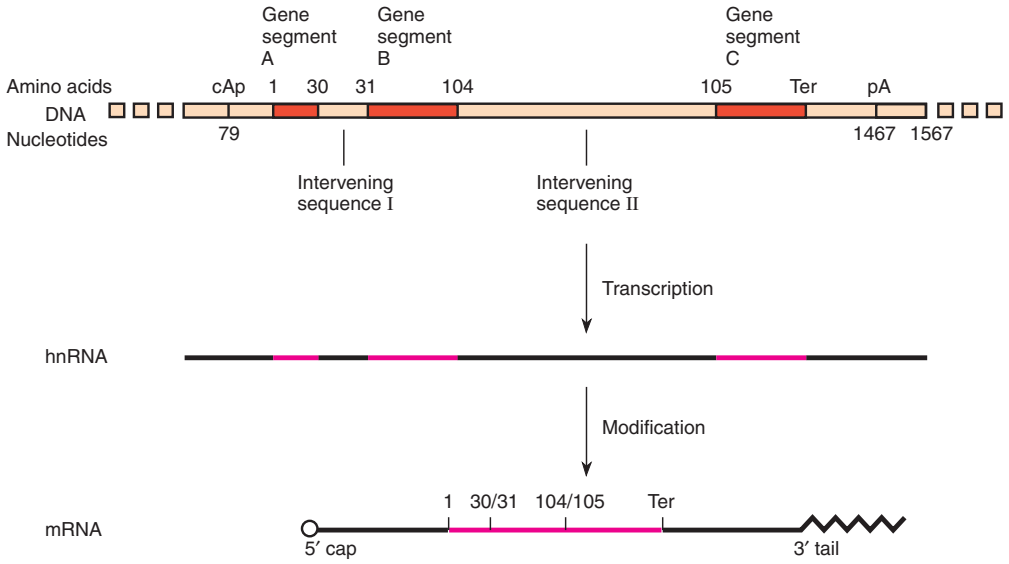
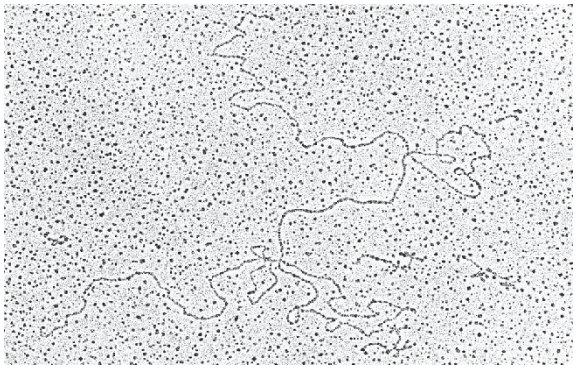
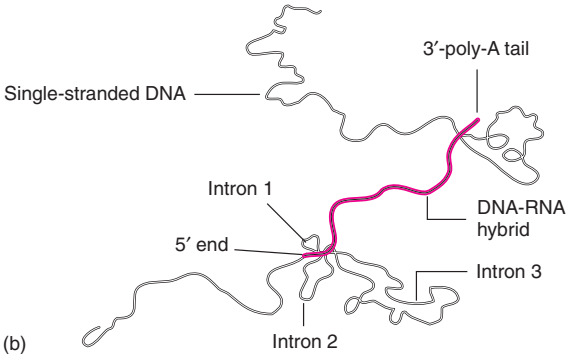


Figure 10.29 Nucleotide sequence of the mouse β-globin major gene. The coding DNA strand is shown; cAp (position 79) indicates the start of the capped mRNA; pA indicates the start of the poly-A tail (position 1467); numbers inside the sequence are adjacent amino acid positions; Ter is the termination codon (position 1334). The three-letter abbreviations (e.g., Met, Val, His) refer to amino acids (see chapter 11). The TATA box begins at position 49. (Source: National Institutes of Health Research by David A. Konkel, et al., "The sequence of the chromosomal mouse β-globin major gene: Homologies in capping, splicing and poly (A) Sites," *Cell,* 15:1125–32, 1978.)



(a)



(b)

Figure 10.30 The mRNA of adenovirus hybridized with its DNA. Three introns are visible as single-stranded DNA loops. They form single-stranded loops because they have nothing in the RNA molecule to hybridize with. Also visible is the poly-A tail of the mRNA. (*a*) Electron micrograph, (*b*) explanatory diagram. ([*a*] Courtesy of Louise T. Chow and Thomas Broker.)

Thomas Cech (1947–). (Courtesy of Dr. Thomas Cech. Photo by Ken Abbott.)

Sidney Altman (1939–). (Courtesy of Dr. Sidney Altman. Photo: Michael Marsland, Yale University Office of Public Affairs.)

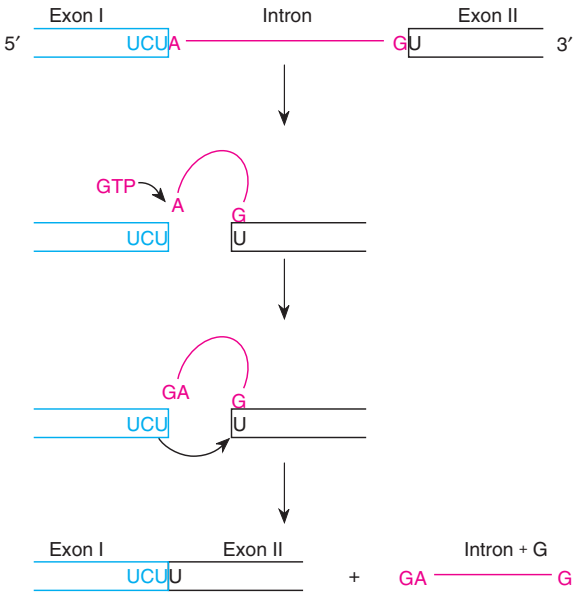


Figure 10.31 Self-splicing of a ribosomal RNA precursor in *Tetrahymena.* An external GTP is required. Two bond transfers produce a shortened RNA and a free intron.

Protein-Mediated Splicing (the Spliceosome)

Eukaryotic nuclear messenger RNAs also have their introns removed by way of a lariat structure, just as in type II introns, but with the help of RNA-protein particles. Figure 10.36 shows consensus sequences in nuclear messenger RNA for the majority of introns. At the left (5′) side of the intron, the GU sequence is invariant, as is the AG at the right (3′) side. The right-most A of the UACUAAC sequence is the branch point of the lariat and is also invariant. (In DNA nucleotides, UACUAAC is TACTAAC; therefore, that region is sometimes referred to as the **TACTAAC box.**)

Unlike the mitochondrial group II introns, however, nuclear messenger RNAs have their introns removed with the help of a protein-RNA complex called a **spliceosome,** named by J. Abelson and E. Brody. The splicing apparatus in eukaryotic messenger RNAs consists of several components called **small nuclear ribonucleoproteins** (discovered and named by J. Steitz and colleagues), abbreviated as **snRNPs** and pronounced "snurps." Five of these particles take part in splicing, each composed of one or more proteins and a small RNA molecule; they are designated U1, U2, U4, U5, and U6. The RNA molecules range in size from 100 to 215 bases. The snRNPs and their associated proteins are located in twenty to forty small regions in the nucleus called *speckles* because of their appearance in the fluorescent microscope.

The RNAs of these particles have been sequenced, and sequencing shows they have regions of complementarity to either sites in the exons, sites in the introns, or sites in the other snRNP RNAs (table 10.5). These sequences, together with the experimental techniques of photocrosslinking and the creation of selective mutations (using techniques of site-directed mutagenesis described in chapter 13) have given us insight into the splicing mechanism. Photocrosslinking tells us which components are in contact. Mutations change pairings of components and may disrupt the structure. The change can be *rescued* (the pairing restored) by making a second change in the complementary RNA. When this happens successfully, the

Joan A. Steitz (1941–).
(Courtesy of Dr. Joan A. Steitz.)

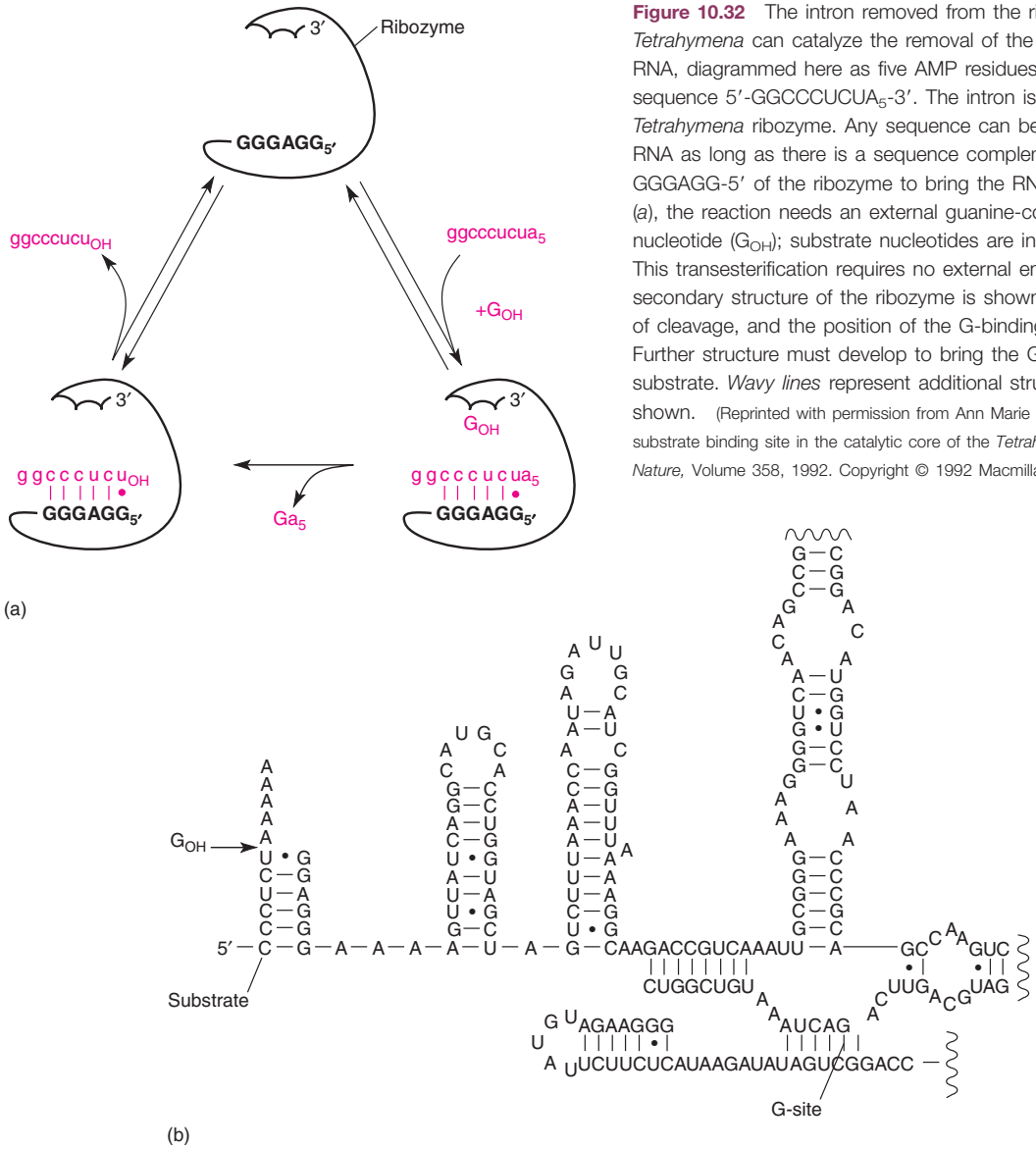


Figure 10.32 The intron removed from the ribosomal RNA of *Tetrahymena* can catalyze the removal of the 3′ end of an RNA, diagrammed here as five AMP residues (a_5) from the sequence 5′-GGCCCUCUA$_5$-3′. The intron is called the *Tetrahymena* ribozyme. Any sequence can be removed from an RNA as long as there is a sequence complementary to the GGGAGG-5′ of the ribozyme to bring the RNA into position. In (*a*), the reaction needs an external guanine-containing nucleotide (G_{OH}); substrate nucleotides are in lowercase letters. This transesterification requires no external energy. In (*b*), the secondary structure of the ribozyme is shown. G_{OH} is the site of cleavage, and the position of the G-binding site is shown. Further structure must develop to bring the G-site to the substrate. *Wavy lines* represent additional structure not shown. (Reprinted with permission from Ann Marie Pyle, et al., "RNA substrate binding site in the catalytic core of the *Tetrahymena* ribozyme," *Nature,* Volume 358, 1992. Copyright © 1992 Macmillan Magazines, Ltd.)

Table 10.5 **The Five Small Nuclear Ribonucleoproteins (snRNPs) Involved in Nuclear Messenger RNA Intron Removal and Their RNAs**

snRNP RNA	Partial Sequence	Complementarity	Role
U1	3′-UCCAUUCAUA	5′ end of intron	Recognizes and binds 5′ site of intron
U2	3′-AUGAUGU	Branch point of intron	Binds branch point of intron
U4	3′-UUGGUCGU . . . AAGGGCACGUAUUCCUU	U6	Binds to (inactivates) U6
U5	3′-CAUUUUCCG	Exon 1 and exon 2	Binds to both exons
U6	3′-CGACUAGU . . . ACA	U2, 5′ site	Displaces U1 and binds 5′ site and U2 at branch point

Source: With permission from the *Annual Review of Genetics*, Volume 28 © 1994 by Annual Reviews www.AnnualReviews.org.

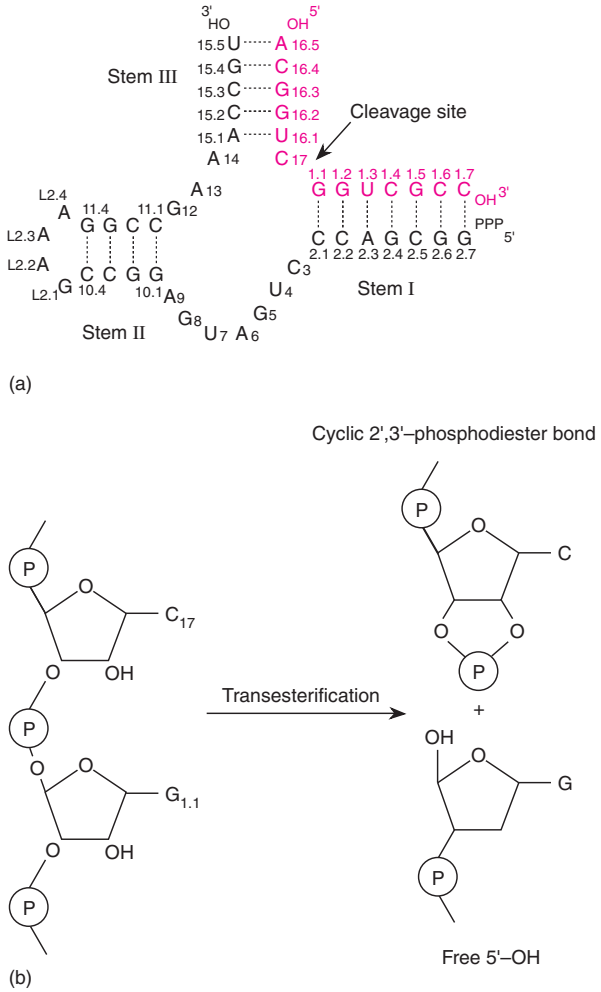


Figure 10.33 The hammerhead ribozyme, first seen in the RNAs of certain viruses (stems I, II, and III). (*a*) The cleavage point of the substrate (*red*) is shown using original sequence numbering, relating to the three stems of the hammerhead-shaped structure. (*b*) The cleavage, a transesterification, creates a cyclic 2′,3′-phosphodiester bond and a free 5′-OH.
(Reprinted with permission from *Nature,* Vol. 372, Heinz W. Pley et al., "Three Dimensional Structure of a Hammerhead Ribozyme." Copyright © 1994 Macmillan Magazines Limited.)

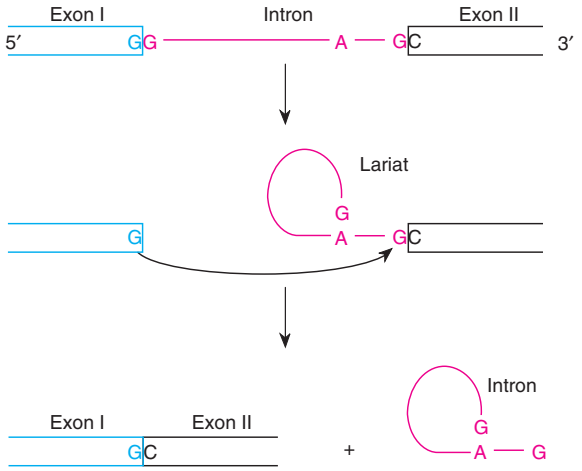


Figure 10.34 Self-splicing of a group II intron results in a lariat configuration of the released intron. No external GTP is required since the first bond transfer takes place with an internal nucleotide, forming the loop of the lariat. A second bond transfer releases the lariat-shaped intron.

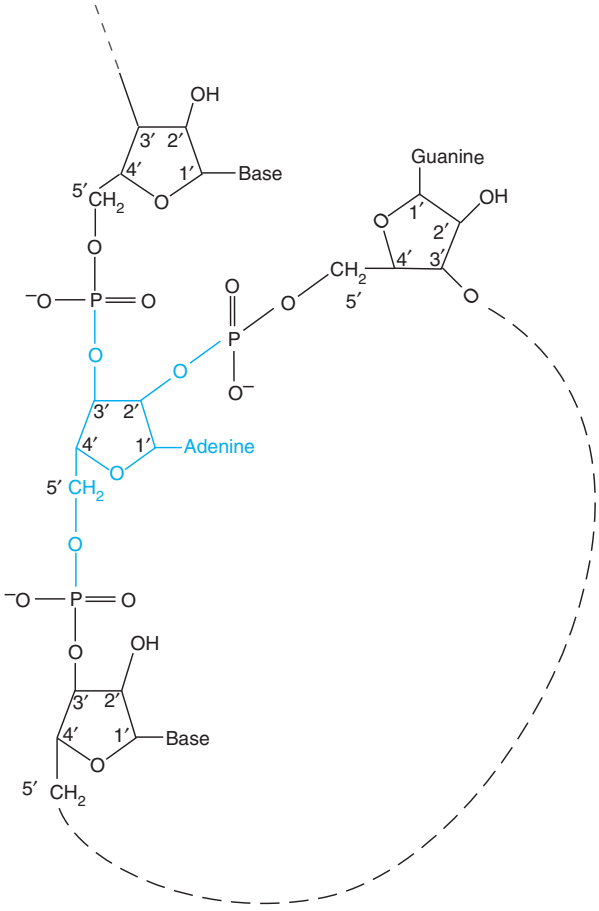


Figure 10.35 The lariat branch point (see fig. 10.34), formed during removal of a group II intron, occurs as three phosphodiester bonds form at the same ribose sugar. (The lariat loop is formed by the 2′-phosphodiester bond.)

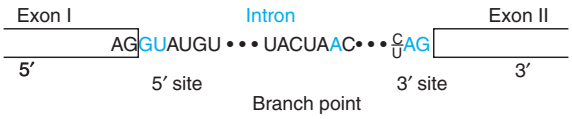


Figure 10.36 Consensus sequences of nuclear introns, showing the 5′ and 3′ sites and the branch point. Letters in *blue* (GU, A, AG) represent invariant bases. The last A (in *blue*) of the UACUAAC sequence is the lariat branch point.

presumed pairing is then confirmed. For example, if an A-U base pair occurs between two pieces of RNA, changing the A to a C disrupts the pairing. However, if the U is converted to a G, the pairing is restored (complementary A-U bases are converted to complementary C-G bases via a noncomplementary C-U intermediate). From these techniques, we believe that the following sequence of events takes place.

First, the U1 snRNP binds at the 5′ site of the intron and the U2 snRNP binds at the branch point (fig. 10.37). The U4, U5, and U6 snRNPs form a single particle. The U4 snRNP releases, freeing the U6 snRNP to bind to the 5′ site, displacing the U1 snRNP. (The U1 snRNP, with the help of other proteins, may bind at the 5′ site simply to mark it and initiate the process.) The U6 snRNP then also binds the U2 snRNP, allowing the lariat to form in the intron. The U5 snRNP binds the two exon ends together, allowing the splice to be completed as the lariat is removed.

The splicing machinery for the majority of introns also includes numerous other polypeptides called *auxiliary* and *splicing factors;* the entire splicing process requires about 50 polypeptides. A second, less common, intron, called the U12-dependent intron, with different consensus sequences, also exists. It is removed by a similar splicing process involving different snRNPs (U11, U12) as well as many components shared with the major spliceosome.

Currently, we believe the splicing out of the intron may be autocatalyzed, just as in the type II self-splicing introns. The spliceosome may have evolved to ensure control over the process, allowing different introns to splice with differing efficiencies and allowing **alternative splicing** to take place. In many eukaryotic genes, alternative paths of splicing can take place—different splice sites may be chosen or splices may be avoided entirely. Thus, a single gene can produce several different proteins, depending on splicing choice. For example, in yeast, the gene *RPL32* codes for a ribosomal protein. When this protein accumulates in excess, it somehow causes intron removal to fail. The result is a nonfunctional messenger RNA and no further RPL32

BOX 10.3

Experimental Methods

Are Viroids Escaped Introns?

Viroids are small (less than four hundred nucleotides), single-stranded RNA circles that act as plant pathogens. They do not have protein coats. In addition, they do not seem to code for any protein. The nature of their pathogenicity is not well understood. In 1986, Gail Dinter-Gottlieb, at the University of Colorado, pointed out numerous regions of homology between viroids and group I introns, supporting proposals by Francis Crick and Theodor Diener that viroids are escaped introns.

Many group I introns form circles after they are released. The self-splicing group I intron of *Tetrahymena thermophila* is 399 bases in circular form, whereas the potato spindle tuber viroid (PSTV) is 359 bases. These similarities of size and shape prompted the search for base homologies. In figure 1, we compare

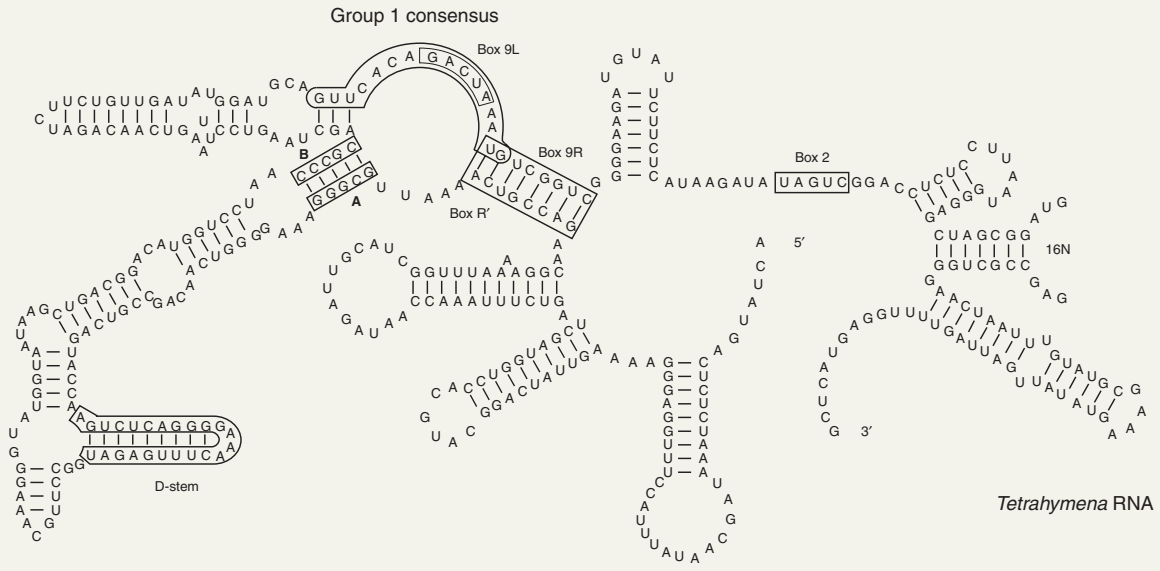


Figure 1 Self-splicing group I *Tetrahymena* intron *(left page)* and potato spindle tuber viroid (PSTV, *right page*); 16N in the *left figure* refers to sixteen nucleotides not shown. Note the similarities around the group 1 consensus area. (From G. Dinter-Gottlieb, *Proceedings of the National Academy of Sciences,* page 6251, 1986.)

protein produced. In human beings, the gene *RBP-MS* can produce at least twelve different transcripts, depending on alternative splicing.

One other mode of protein-mediated intron removal is known. Nuclear transfer RNAs have introns that are not self-splicing but are removed by an endonuclease; the exons are subsequently joined by a ligase. Archaean bacteria seem to have this type of intron.

Intron Function and Evolution

Since the discovery of introns, geneticists have been trying to figure out why they exist. Several views have arisen. Walter Gilbert suggested that introns separate exons (coding regions) into functional domains—that is, they separate different exons that presumably have specific tasks. In a given protein, one exon might code for a membrane-binding region, one might code for the active site of the enzyme, and one might code for ATPase activity. By recombinational mechanisms, or by excluding an exon during intron removal, **exon shuffling** would allow the rapid evolution of new proteins whose structures would be conglomerates of various functional domains. In a 1990 article in *Science,* Gilbert, with two colleagues, calculated that all proteins in eukaryotes can be accounted for by as few as one thousand to seven thousand

the *Tetrahymena* intron with PSTV. Note that both have an extensive secondary structure (stem-loops) and similarities of some sequences. Most notable is the box 9L similarity. This box lies within a 16-base consensus sequence of all group I introns and has similarities to the corresponding sequence in PSTV. Note the general shape around the group I consensus region: two stems to the left and one to the right with some homologies. Note also the D-stem similarities.

These similarities strongly indicate that viroids and group I introns are related. Whether viroids are escaped introns or both evolved from a common ancestor has, as yet, not been resolved.

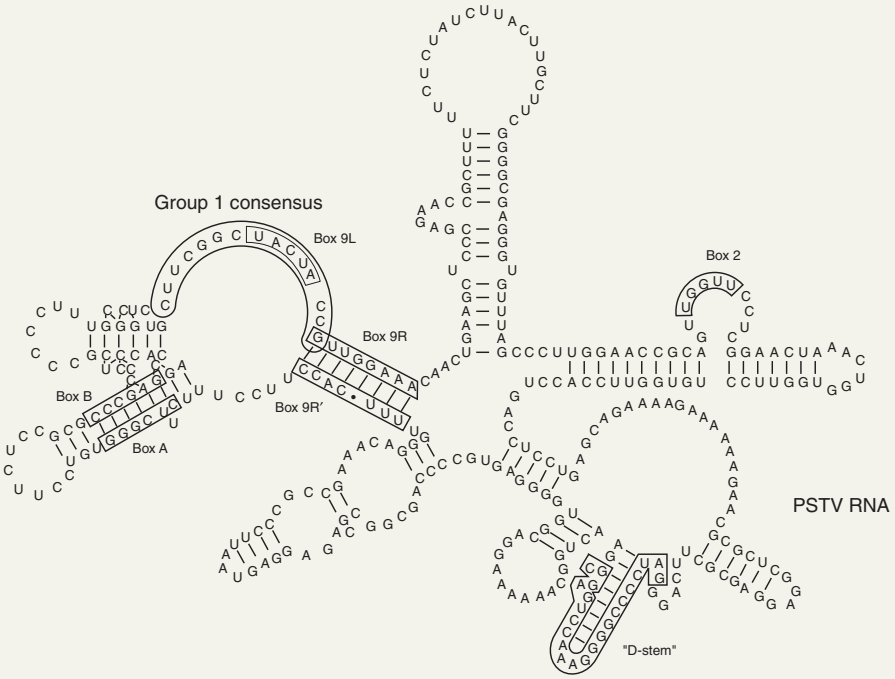


Figure 1 (continued)

exons; all proteins may be conglomerates of this primordial number. However, this view is controversial.

J. Darnell and W. F. Doolittle have expanded Gilbert's idea of exon shuffling into the *introns-early* view. They suggest that introns arose before the first cells evolved. After eukaryotes evolved from prokaryotes, the prokaryotes lost their introns. This is supported by the evidence that, generally, prokaryotes lack introns. This view is also consistent with the opinion that the original genetic material was RNA. In this "RNA world," introns arose as part of the genetic apparatus; they were the first enzymes (ribozymes).

An alternative view is that introns arose later in evolution, after the eukaryotes split from the prokaryotes. At first, the justification for this *introns-late* view was that introns evolved late to give the organism the ability to evolve quickly to new environments by an exon-shuffling type of mechanism. However, evolutionary biologists don't accept the rationale of evolution based on future needs. An alternative explanation is that introns are actually invading "selfish DNA," DNA that can move from place to place in the genome without necessarily providing any advantage to the host organism. We call these "jumping genes" transposons and discuss them at length in chapters 14 and 16. Thus, both time frames for the development of introns—late or early—have conceptual support.

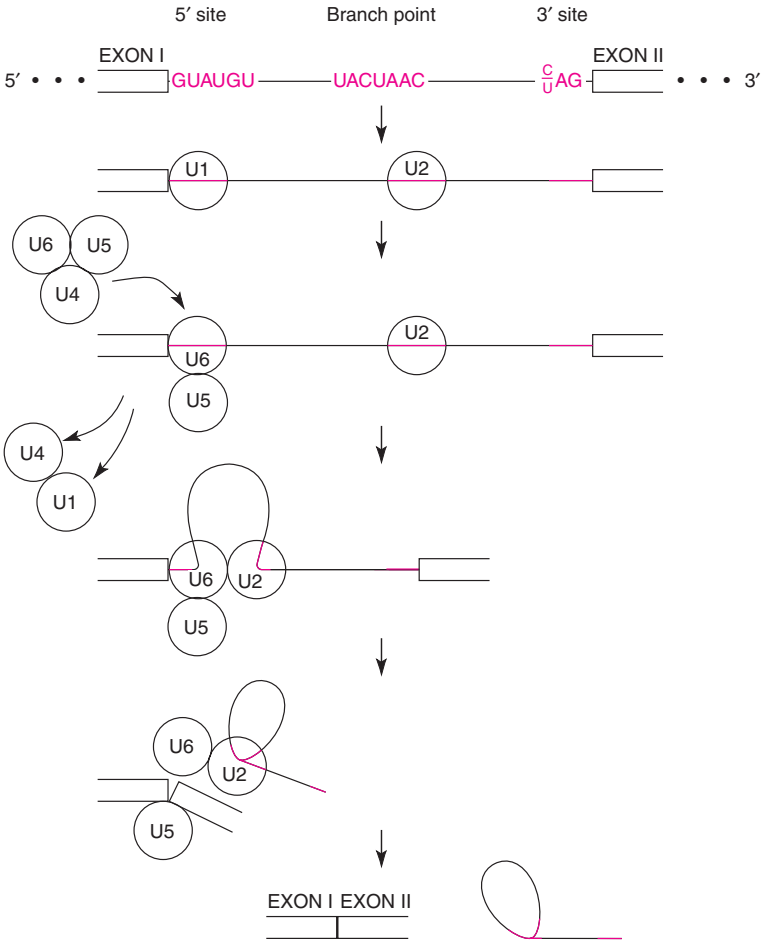


Figure 10.37 Sequence of steps, explained in the text, in which U1, U2, U4, U5, and U6 snRNPs take part in intron removal in a nuclear RNA.

Evidence exists to support both views. Gilbert's exon-shuffling view is supported by the analysis of some genes that do indeed fit the pattern of exons coding for functional domains of a protein. (Analysis consists of DNA sequencing, RNA sequencing, and protein structural analysis.) For example, the second of three exons of the globin gene binds heme. Similarly, the human low-density lipoprotein receptor is a mosaic of exon-encoded modules shared with several other proteins. Autocatalytic properties of introns lend credence to the view that RNA was the original genetic material and that introns can move within a genome.

Additional evidence for the introns-early hypothesis includes the discovery of several introns in phage genes and introns in transfer RNA and ribosomal RNA genes in ancient bacteria (archaebacteria). Until recently, however, no introns were known in the true bacteria (eubacteria). That changed with recent work from the labs of D. Shub and J. Palmer, who independently discovered an intron in a transfer RNA gene in seven species of cyanobacteria (blue-green algae of the eubacteria). This intron was suspected to exist because it occurred in the equivalent chloroplast gene; the chloroplast evolved from an invading cyanobacterium. However, this discovery has been viewed as supporting both the introns-early and introns-late view. The introns-early supporters say this evidence confirms that introns arose before the eukaryotes-prokaryotes split. Introns-late supporters say they expect to see some introns in prokaryotes because of the mobility these bits of genetic material have.

Both the introns-early and the introns-late views may be correct. It is possible that introns arose early, were lost by the prokaryotes, which prioritized small genomes and rapid, efficient DNA replication, and later evolved to produce exon shuffling in eukaryotes.

RNA Editing

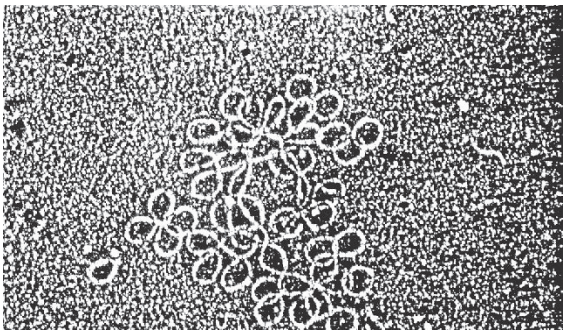
In the last few years, several examples have arisen in which DNA sequence does not predict protein sequence. In several cases, changes in the protein occur that could have only come about by inserting or deleting nucleotides in the messenger RNA before it is translated. This insertion or deletion is almost exclusively of uridines. The process is termed **RNA editing.**

RNA editing was particularly evident in the mitochondrial proteins of a group of parasites, the trypanosomes (some of which cause African sleeping sickness); in one case, more than 50% of the nucleotides in the messenger RNA were added uridines. Uridines were also deleted from the original sequence. These parasites had another mysterious trait—the existence of minicircles and maxicircles of DNA in specialized mitochondria called *kinetoplasts.* In the average kinetoplast, there are about fifty maxicircles and about five thousand minicircles, concatenated like chain links (fig. 10.38*a*). The maxicircles contain genes for mitochondrial function (see chapter 17); as L. Simpson and his colleagues showed in 1990, both maxicircles and minicircles are templates for **guide RNA (gRNA),** RNA that guides the process of messenger editing.

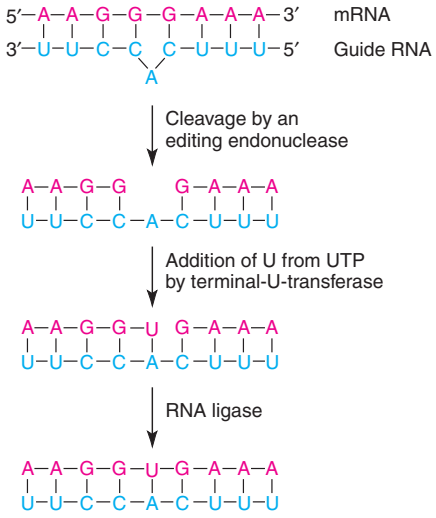
The guide RNA forms a complement with the messenger RNA to be edited; however, the guide RNA has the sequence complementary to that of the *final* messenger RNA, the one with bases added. Since the bases have not yet been added, a bulge occurs in the guide RNA where the complement to be added is (fig. 10.38*b*). The messenger RNA is then cleaved opposite the bulge by an editing endonuclease. A uridylate (U) is brought into the messenger RNA as a complement to the adenine (A) with the enzyme terminal-U-transferase. An RNA ligase then closes the nick in the messenger RNA, which now has a uridylate added.

An exciting outcome of this research, aside from learning about a novel mechanism of messenger RNA processing, is the possibility of clinical rewards. Anytime there is a specialized pathway in a parasite not found in its host, it is possible to use that pathway to attack the parasite. Thus, this research might lead to new ways of combating these trypanosome parasites.

RNA editing also occurs in other species and by different mechanisms. For example, in the apolipoprotein-B (*apoB*) gene in mammals, one gene produces two forms of the protein. In one case, nucleotide 6666, a cytosine, is modified by deamination to a uracil in the messenger RNA, resulting in the termination of translation and a protein about half the normal size. RNA editing also occurs in plant mitochondria and chloroplasts in which the usual change is also a cytosine to a uracil. RNA editing is thus routinely seen in specific examples of posttranscriptional RNA modification in both animals and plants.



(a)



(b)

Figure 10.38 RNA editing. (*a*) Eight hundred seventy base pair minicircles of DNA from *Leishmania tarentolae.* (*b*) Mechanism by which a guide RNA is involved in the editing of a messenger RNA. After the cycle shown, a uridine-containing nucleotide has been added to the messenger RNA. The guide RNA has the sequence complementary to the messenger RNA with the base already added. ([*a*] Courtesy of Larry Simpson.)

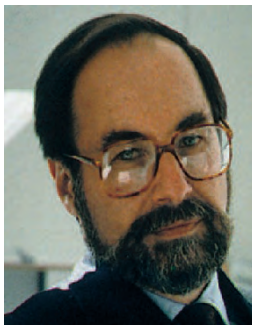
UPDATED INFORMATION ABOUT THE FLOW OF GENETIC INFORMATION

The original description of the central dogma included three information transfers that were presumed to occur even though they had not been observed (see fig. 10.1). Since then, researchers have documented these three transfers: reverse transcription, RNA self-replication, and the direct involvement of DNA in translation (fig. 10.39).

Reverse Transcription

First, the return arrow from RNA to DNA in figure 10.39 indicates that RNA can be a template for DNA synthesis. All RNA tumor viruses, such as Rous sarcoma virus, as well as the AIDS virus, can make an RNA-dependent DNA polymerase (often referred to as **reverse transcriptase**) that synthesizes a DNA strand complementary to the viral RNA. (H. Temin and D. Baltimore received Nobel prizes for their discovery of this polymerase enzyme.) This enzyme is involved in a tumor virus's infection of a normal cell and the transformation of that cell into a cancerous cell. When the viral RNA enters a cell, it brings reverse transcriptase with it. The enzyme synthesizes a DNA-RNA double helix, which then is enzymatically converted into a DNA-DNA double helix that can integrate into the host chromosome. After integration, the DNA is transcribed into copies of the viral RNA, which are both translated and packaged into new viral particles that are released from the cell to repeat the infection process. (We cover this material in more detail in chapters 13 and 16.)

Howard Temin (1934–1994). (Courtesy of Dr. Howard Temin. UW photo media.)



David Baltimore (1938–). (Courtesy of Kucerea and Company/Laxenburger Strasse 58.)

RNA Self-Replication

The second modification to the original central dogma is the verification that RNA can act as a template for its own replication. This process has been observed in a small class of phages. These **RNA phages,** such as R17, f2, MS2, and Qβ, are the simplest phages known. MS2 contains about thirty-five hundred nucleotides and codes for only three proteins: a coat protein, an attachment protein (responsible for attachment to and subsequent penetration of the host), and a subunit of the enzyme **RNA replicase.** The RNA replicase subunit combines with three of the cell's proteins to form RNA replicase, allowing the single-stranded RNA of the phage to replicate itself.

Since the new protein needed to construct the RNA replicase enzyme must be synthesized before the phage can replicate its own RNA, the phage RNA must first act as a messenger when it infects the cell. Thus, protein synthesis is taking place without a preceding transcription process. The viral genetic material, its RNA, is first used as a messenger in the process of translation and then used as a template for RNA replication.

DNA Involvement in Translation

In the mid-1960s, B. J. McCarthy and J. J. Holland showed that under certain experimental conditions, denatured (single-stranded) DNA could bind to ribosomes and be translated into proteins. The experimental conditions usually involved the addition of antibiotics that interacted with the DNA or the ribosome. Direct translation of DNA is not known to occur naturally.

Even in our updated central dogma (fig. 10.39), no arrows originate at the protein. In other words, protein cannot self-replicate, nor can it use amino acid sequence information to reconstruct RNA or DNA. Crick has called these arrows "forbidden transfers." We know of no cellular machinery that can produce these forbidden processes. In the next chapter, we continue this discussion of protein synthesis by describing the process of translation, in which the information in messenger RNA is used to form the sequences of amino acids in proteins.

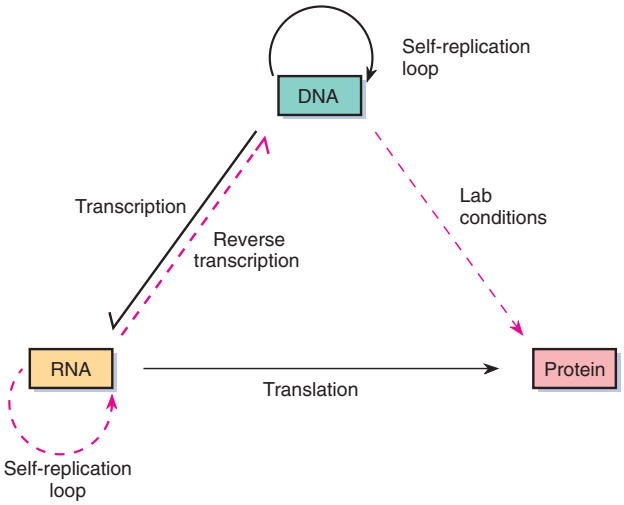


Figure 10.39 An updated version of Crick's central dogma, showing all known paths of genetic information transfer. Paths confirmed since Crick proposed the original central dogma appear as dashed *red* lines (reverse transcription, RNA self-replication, and direct DNA translation). Direct DNA translation is known only under laboratory conditions: the process apparently does not occur naturally. There is no known information flow beginning with protein.

SUMMARY

The central dogma is a description of how genetic information is transferred among DNA, RNA, and protein. In chapter 9, we described the DNA self-replication loop. In this chapter, we described the transcriptional process, in which DNA acts as a template for the production of RNA.

STUDY OBJECTIVE 1: To examine the types of RNA and their roles in gene expression 245–246, 256–260

Messenger RNA (mRNA) is a complementary copy of the DNA of a gene that carries the information of the gene to the ribosomes, where protein synthesis actually takes place. Transfer RNAs (tRNAs) transport the amino acid building blocks of proteins to the ribosome. Complementarity between the messenger RNA codon and the transfer RNA anticodon establishes the amino acid sequence in the synthesized protein ultimately specified by the gene. Ribosomal RNA (rRNA) is also involved in this process of gene-directed protein synthesis.

STUDY OBJECTIVE 2: To look at the process of transcription, including start and stop signals, in both prokaryotes and eukaryotes 246–256

Intracellular RNA is single-stranded, although extensive intramolecular stem-loop structures may form. At any one gene, RNA is transcribed from only one strand of the DNA double helix. The transcribing enzyme is RNA polymerase. In *E. coli,* the core enzyme, when associated with a sigma factor, becomes the holoenzyme that recognizes the transcription start signals in the promoter. Several consensus sequences define a promoter. In prokaryotes, termination of transcription requires a sequence on the DNA, called the terminator, that causes a stem-loop structure to form in the RNA. Sometimes the rho protein is required for termination (in rho-dependent, as compared with rho-independent, termination). In eukaryotes, there are three RNA polymerases. Eukaryotic genes have promoters with sequences analogous to those in prokaryotic promoters as well as enhancers that work at a distance.

The ribosome is made of two subunits, each with protein and RNA components. Transfer RNAs are charged with their particular amino acids by enzymes called aminoacyl-tRNA synthetases. Each transfer RNA has about eighty nucleotides, including several unusual bases. All transfer RNAs have similar structures and dimensions. Transfer RNAs and ribosomal RNAs are modified from their primary transcripts.

STUDY OBJECTIVE 3: To investigate posttranscriptional changes in eukaryotic messenger RNAs, including an analysis of intron removal 260–276

Prokaryotic messenger RNAs are transcribed with a leader before, and a trailer after, the translatable part of the gene. In prokaryotes, translation begins before transcription is completed. In eukaryotes, these processes are completely uncoupled—transcription is nuclear and translation is cytoplasmic. Eukaryotic messenger RNA is modified after transcription: a cap and tail are added, and intervening sequences (introns) are removed, before transport into the cytoplasm. Introns can be removed by self-splicing or with the aid of the spliceosome, composed of small nuclear ribonucleoproteins (snRNPs). It is not known whether introns arose early or late in evolution or what their functions are. In some organisms, such as trypanosomes, RNAs can be edited further by the addition or deletion of nucleotides under the direction of guide RNA.

The study of several RNA viruses has shown that RNA can act as a template to replicate itself and to synthesize DNA; under laboratory conditions, DNA can be translated directly into protein. These discoveries add new directions of information transfer to the central dogma.

SOLVED PROBLEMS

PROBLEM 1: What would be the sequence of segments on a prokaryotic messenger RNA with more than one gene present?

Answer: The transcript would have unmodified 5′ (leader) and 3′ (trailer) ends. Reading the sequence of nucleotides on the RNA, you would come across an initiation codon (AUG) and then, after perhaps nine hundred more nucleotides, a termination codon (UAA, UAG, or UGA). The nine hundred nucleotides would be those translated into the protein. Then there would be a spacer region of nucleotides, followed by another initiation codon, intervening nucleotides that are translated into amino acids, and a termination codon. This sequence of initiation codon, codons to be translated, a termination codon, and spacer RNA would be repeated for as many genes as are present in the messenger RNA.

PROBLEM 2: Can one nucleotide be a conserved sequence?

Answer: Conserved sequences are invariant sequences of DNA or RNA recognizable to either a protein or a

complementary sequence of DNA or RNA. However, in group II introns, an adenine is needed near the 3′ end of the intron for lariat formation. Thus, this single nucleotide, given its relative position in the intron and possible surrounding bases, is a conserved sequence of one.

PROBLEM 3: Why might *E. coli* not have a nucleolus?

Answer: The nucleolus is the site of ribosomal construction in eukaryotes. It is centered at the nucleolus organizer, the tandemly repeated gene coding for the three larger pieces of ribosomal RNA. In *E. coli,* there are only five to ten copies of the ribosomal RNA gene, whereas there is usually an order of magnitude or more copies in eukaryotes. Thus, the simplest reason that a nucleolus is not visible in *E. coli* is because there are too few copies of the gene around which a nucleolus forms.

PROBLEM 4: If this sequence of bases represents the start of a gene on double-stranded DNA, what is the sequence of the transcribed RNA, what is its polarity, and what is the polarity of the DNA?

G C T A C G G A T T G C T G
C G A T G C C T A A C G A C

Answer: Begin by writing the complementary strand to each DNA strand: C G A U G C C U A A C G A C for the top, and G C U A C G G A U U G C U G for the bottom. Now look for the start codon, AUG. It is present only in the RNA made from the top strand, so the top strand must have been transcribed. The polarity of the start codon is 5′-A U G-3′. Since transcription occurs 5′ → 3′, and since nucleic acids are antiparallel, the left end of the top DNA strand is the 3′ end.

EXERCISES AND PROBLEMS*

TYPES OF RNA

1. Diagram the relationships of the three types of RNA at a ribosome. Which relationships make use of complementarity?

PROKARYOTIC DNA TRANSCRIPTION

2. How could DNA-DNA or DNA-RNA hybridization be used as a tool to construct a phylogenetic (evolutionary) tree of organisms?
3. Assume that prokaryotic RNA polymerase does not proofread. Do you expect high or low levels of error in transcription as compared with DNA replication? Why is it more important for DNA polymerase than RNA polymerase to proofread?
4. What are the transcription start and stop signals in eukaryotes and prokaryotes? How are they recognized? Can a transcriptional unit include more than one translational unit (gene)? (*See also* EUKARYOTIC DNA TRANSCRIPTION)
5. What is a consensus sequence? a conserved sequence?
6. What would the effect be on transcription if a prokaryotic cell had no sigma factors? no rho protein?
7. Draw a double helical section of prokaryotic DNA containing transcription start and stop information. Give the base sequence of the messenger RNA transcript.
8. In what ways does the transcriptional process differ in eukaryotes and prokaryotes? (*See also* EUKARYOTIC DNA TRANSCRIPTION)
9. What is a stem-loop structure? an inverted repeat? a tandem repeat? Draw a section of a DNA double helix with an inverted repeat of seven base pairs.
10. What is the function of each of the following sequences: TATAAT, TTGACA, TATA, TACTAAC? What is a Pribnow box? a Hogness box? (*See also* EUKARYOTIC DNA TRANSCRIPTION)
11. What is footprinting? How did it help define promoter sequences?
12. What are the differences between rho-dependent and rho-independent termination of transcription?
13. What are the differences between a σ^{70} and a σ^{32}?
14. Draw a typical mature messenger RNA molecule of a prokaryote and a eukaryote. Label all regions. (*See also* EUKARYOTIC DNA TRANSCRIPTION)
15. Determine the sequence of both strands of the DNA from which this RNA was transcribed. Indicate the 5′ and 3′ ends of the DNA and, with an arrow, which strand was transcribed.

5′-C C A U C A U G A C A G A C C C U U G C U A A C G C-3′

16. The following DNA fragment was isolated from the beginning of a gene. Determine which strand is transcribed, indicate the polarity of the two DNA strands, and then give the sequence of bases in the resultant messenger RNA and its polarity.

C C C T A C G C C T T T C A G G T T
G G G A T G C G G G A A A G T C C A A

*Answers to selected exercises and problems are on page A-11.

17. The following DNA fragment represents the beginning of a gene. Determine which strand is transcribed, and indicate polarity of both strands in the DNA.

A T G A T T T A C A T C T A C A T T T A C A T T
T A C T A A A T G T A G A T G T A A A T G T A A

18. The following sequence of bases in a DNA molecule is transcribed into RNA:

C C A G G T A T A A T G C T C C A G T A T G G C A T G G T A C T T C C G G
↑

If the T (*arrow*) is the first base transcribed, determine the sequence and polarity of bases in the RNA, and identify the Pribnow box and the initiator codon.

19. You have isolated a mutant that makes a temperature-sensitive rho molecule; rho functions normally at 30° C, but not at 40° C. If you grow this strain at both temperatures for a short period of time and isolate the newly synthesized RNA, what relative size RNA do you expect to find in each case?

20. Suppose you repeat the experiment in problem 19 and find the same size RNA made at both temperatures. Provide two possible explanations for this unexpected finding.

21. Why do you think most promoter regions are A-T rich? (*See also* EUKARYOTIC DNA TRANSCRIPTION)

EUKARYOTIC DNA TRANSCRIPTION

22. Would introns be more or less likely than exons to accumulate mutations through evolutionary time?

23. What would be the effect on the final protein product if an intervening sequence were removed with an extra base? one base too few?

24. What is heterogeneous nuclear messenger RNA? What are small nuclear ribonucleoproteins?

25. What product would DNA-RNA hybridization produce in a gene with five introns? no introns? Draw these hybrid molecules.

26. What are the recognition signals within the majority of introns?

27. What are the differences between group I and group II introns?

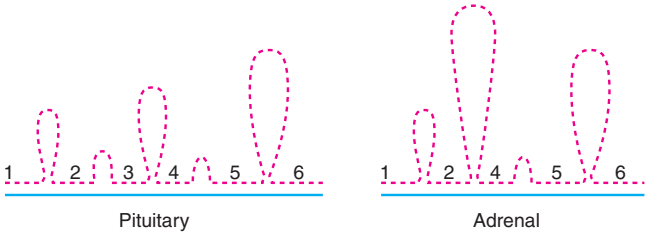
28. Diagram ribozyme functioning in a group I intron.

29. How does a spliceosome work? What are its component parts?

30. What is a transcriptional factor? an enhancer?

31. In the following drawing of a eukaryotic gene, solid red lines represent coding regions, and dashed blue lines represent introns. Draw the RNA-DNA hybrid that would result if cytoplasmic messenger RNA is hybridized to nuclear DNA.

32. RNA-DNA hybrids are formed by using messenger RNA for a given gene that is expressed in the pituitary and the adrenal glands. The DNA used in each case is the full-length gene. Based on the figure, provide an explanation for the different hybrid molecules. DNA is a dashed red line; RNA is a solid blue line.



33. Enhancers can often exert their effect from a distance; some enhancers are located thousands of bases upstream from the promoter. Propose an explanation to account for this.

UPDATED INFORMATION ABOUT THE FLOW OF GENETIC INFORMATION

34. How do prions relate to the central dogma of figure 10.39?

CRITICAL THINKING QUESTIONS

1. What are the upper limits to the size of a gene in eukaryotes?

2. Present a scenario of an activator controlling transcription in eukaryotes.

Suggested Readings for chapter 10 are on page B-7.