The deadly nightshade (Amanita phalloides), one of the most poisonous mushrooms known, contains $\alpha$-amanitin, which blocks eukaryotic messenger RNA synthesis because of its potent inhibition of RNA polymerase II.
Transcription and Post-transcriptional Processing

We turn now to transcription, the process by which information stored in the nucleotide sequence of DNA is read out through the template-dependent synthesis of polyribonucleotides. Mechanistically, transcription is similar to DNA replication, particularly in the use of nucleoside triphosphate substrates and the template-directed growth of nucleic acid chains in a 5′ → 3′ direction. There are, however, two major differences. First, with a few exceptions, only one strand of the DNA duplex is used as a template and transcribed for any particular gene, and second, only part of the entire genome is transcribed in any one cell. The importance of understanding transcription lies in regulation—the processes used to select particular genes and template strands for transcription. This selection, in large part, governs the metabolic capabilities of a cell. These regulatory mechanisms chiefly operate at the levels of initiation and termination of transcription, through the actions of proteins that contact DNA in a highly site-specific manner. Figure 24.1 gives a preview of what we have learned about transcriptional regulation through the analysis of DNA–protein interactions. We discuss this topic both here and in Chapter 26.
The products of transcripive RNA synthesis are rarely used directly. In eukaryotic cells transcription products are subject to further processing, including trimming, cutting, splicing, and modifications at the 3’ and 5’ ends. Comparable processing occurs for prokaryotic ribosomal and transfer RNAs. Therefore, we also discuss post-transcriptional processing in this chapter. We are concerned primarily with the synthesis and processing of messenger RNA, ribosomal RNA, and transfer RNA, the species most directly involved in gene expression. However, some RNA transcripts serve additional functions, including catalysis (Chapters 8 and 25) and the telomerase reaction (Chapter 22). In Chapter 26, we will discuss other RNA species and functions—RNA editing, riboswitches, small interfering RNAs, microRNAs, and noncoding RNAs.

### 24.1 DNA as the Template for RNA Synthesis

The idea that RNA molecules are complementary copies of DNA templates originated from genetic analysis. In France’s Pasteur Institute, François Jacob, Jacques Monod, and André Lwoff were studying genes controlling the utilization of lactose in *E. coli* and the control of lysogeny in bacteriophage λ. Recall from Chapter 23 that lysogeny is the state of gene repression after insertion of a circular bacteriophage genome into the chromosome of a host bacterium. Both the lactose and phage λ systems are described more fully in Chapter 26. Here we focus upon lactose utilization.

**CONCEPT** Our original concepts of RNA synthesis came from genetic studies that predicted the existence of messenger RNA.

### The Predicted Existence of Messenger RNA

Lactose utilization in *E. coli* was known to be controlled by three enzymes, whose genes are adjacent on the chromosome. One of these enzymes is β-galactosidase, which hydrolyzes lactose and other β-galactosides. When bacteria are grown in a medium containing both glucose and lactose, the genes controlling lactose utilization are silent, and glucose is preferentially metabolized. Once the glucose supply is exhausted, the three enzymes of lactose utilization are rapidly synthesized, allowing the bacteria to take up and metabolize lactose. This enzyme **induction** is rapidly reversed upon adding glucose back to the medium, whereupon the β-galactosidase activity reverts to a low level. The rapid changes in β-galactosidase-forming capacity suggested that the template for synthesizing this enzyme is metabolically unstable—synthesized rapidly on demand and degraded when the stimulus to induction is removed. Ribosomal and transfer RNAs are metabolically stable. Hence, these species were unlikely to be intermediates in information transfer. What might these observations tell us about RNA synthesis?

Jacob and Monod analyzed *E. coli* mutants that displayed faulty control over induction of the lactose-utilizing enzymes. Some expressed all three genes at high levels even when lactose or a similar inducer was absent, and others could not produce any of the enzymes, even after the addition of lactose to a bacterial culture. Based on these studies, as well as parallel work by Lwoff with phage λ, in 1961 Jacob and Monod proposed a unifying hypothesis of gene regulation in which transcription, or copying information encoded in...
DNA, was regulated specifically at the level of initiation. Hypothetical regulatory elements called repressors and operators controlled the synthesis of other hypothetical entities called messenger RNAs (mRNAs). mRNA was postulated to be a complementary copy of the DNA that encompassed a set of structural genes, which encode proteins, as schematized in Figure 24.2. A set of contiguous genes plus adjacent regulatory elements that control their expression was termed an operon, and the Jacob–Monod hypothesis thus became known as the operon model.

Jacob and Monod predicted several characteristics of the hypothetical mRNA. First, they predicted a high rate of mRNA synthesis followed by rapid degradation, which would explain the fast turn-on of the genes after induction by lactose and turn-off in the presence of another preferentially used sugar. Second, because of rapid synthesis and degradation, they expected mRNA to accumulate rapidly but not to high steady-state levels. Third, because they thought that the messenger was a copy of two or more contiguous genes, they expected it to be fairly large and part of a heterogeneous size class of RNA molecules. Finally, if the mRNA was a complementary copy of DNA, its nucleotide sequence should be identical to a portion of one of the DNA strands.

### CONCEPT: Bacterial genetics

Predicted messenger RNA to be a collection of metabolically active RNAs that are present in low abundance, heterogeneous in size, and complementary in sequence to portions of the DNA genome.

### T2 Bacteriophage and the Demonstration of Messenger RNA

The first physical demonstration of mRNA came from work with the closely related T2 and T4 bacteriophages. Infection by these large DNA viruses arrests all expression of host-cell genes, and no significant accumulation of RNA can be detected after infection. However, in 1956 the use of radioisotopes led to detection of a distinctive RNA in T2-infected E. coli. When infected cultures were labeled for short periods (a 3- or 4-minute "pulse") with $[^{32}\text{P}]$orthophosphate, about 2% of the total RNA became radioactive. This radiolabeled RNA had two properties that led to its eventual identification as viral mRNA. First, it was metabolically unstable; after termination of the labeling interval, this RNA rapidly lost radioactivity, as if the labeled species were being degraded. Second, the RNA seemed to be a product of viral DNA metabolism because its nucleotide composition resembled that of T2 DNA; the radiolabeled RNA was rich in adenine and uracil and low in guanine and cytosine; T2 DNA has a high A + T content.

Additional evidence came from sucrose gradient centrifugation of the rapidly labeled phage RNA, which showed that the labeled material sediments heterogeneously and distinctly from any of the known rRNA or tRNA species. In the experiment of Figure 24.3(a), an ultracentrifuge tube contains buffer with a linear gradient of sucrose concentration. The RNA solution being analyzed is carefully layered onto the top of the gradient. The sucrose minimizes convection, so that in a centrifugal field molecules sediment as discrete bands, at rates related to their size and shape. This differs from equilibrium gradient centrifugation, as in the Meselson–Stahl experiment (Chapter 4), where molecules are separated on the basis of density differences (mass per unit volume).

In the experiment of Figure 24.3(a), fractions obtained after centrifugation were analyzed for their absorbance profile, which measures RNA concentration, and radioactivity, which measures the relative amount of RNA synthesized during the pulse. The UV absorbance profile shows two large peaks at 16S and 23S, and a smaller peak at 4S. (S stands for “Svedberg units,” a measure of sedimentation rate; it is related to molecular size and shape.) The 16S and 23S peaks represent rRNA, and the 4S peak is mostly tRNA, which is a much smaller molecule. The radioactivity was distributed heterogeneously throughout the gradient, suggesting that the labeled RNA is a collection of molecules of varying sizes, none of which was present in great abundance.

Benjamin Hall and Sol Spiegelman established that this phage RNA is a viral gene product when they carried out the first DNA–RNA hybridization experiment, showing that the labeled RNA was complementary in sequence to phage DNA. DNA and RNA were heated and slow-cooled together, so that RNA could base-pair with DNA of complementary nucleotide sequence. The earliest experiments were based on the fact that RNA is denser than DNA (why is this so?). Thus, a DNA–RNA hybrid could be detected in equilibrium gradient centrifugation as a species of intermediate density containing label derived from both DNA and RNA. Such a hybrid was formed when
T2 RNA was heated and slowly cooled along with T2 DNA, but not when the DNA came from *E. coli*. Additional density-labeling experiments established that phage proteins were synthesized on ribosomes that had been formed before infection. These experiments showed the ribosome to be a nonspecific workbench on which any protein could be assembled, depending on which template became associated with that workbench.

**RNA Dynamics in Uninfected Cells**

The experiments described in the previous section supported the existence of mRNA in phage-infected *E. coli*. What about uninfected bacteria? Spiegelman and his colleagues showed that pulse-labeled RNA from uninfected *E. coli* hybridized to *E. coli* DNA. At very short labeling intervals, the sedimentation pattern showed incorporation into both rRNA and tRNA species and a heterogeneously sedimenting species (Figure 24.3(c)). After a “chase”, when the radiolabeled phosphate was removed, the radioactivity profile followed the absorbance profile, showing that all RNA species had become labeled to equivalent specific radioactivities. This finding is consistent with the postulated short lifetime of mRNA (t½ = 2–3 minutes). mRNA would reach its maximal radioactivity within just a few minutes, but during the chase, mRNA turnover would release nucleotides that could flow into stable RNA molecules. Because those stable RNA molecules do not turn over, label accumulates, and the fraction of total label in the stable RNA species continues to increase. Consistent with this idea, Spiegelman also showed that highly labeled ribosomal and transfer RNAs hybridize to *E. coli* DNA, demonstrating that all three major classes of RNA are synthesized from template DNA strands.

As noted previously, the earliest DNA–RNA hybridization experiments involved gradient centrifugation, a laborious technique. Spiegelman and his colleagues made the important discovery that single-stranded DNA binds irreversibly to membrane filters made of material such as nitrocellulose. This technique allowed rapid hybridization analysis of a large number of samples because a radiolabeled RNA could hybridize to denatured DNA immobilized on a nitrocellulose filter. After washing unbound...
radioactivity off the filter, the extent of hybridization could be determined simply by counting the radioactivity of the filter. The same principle—immobilization of DNA on nitrocellulose followed by the analysis of bound radioactivity—underlies Southern blotting (described in Chapter 21), which is now more widely used to analyze gene organization and expression. These developments also led to microarray technology, in which thousands of DNA–RNA hybridization reactions are analyzed on a single DNA chip (see Tools of Biochemistry 24A; see also Figure 1.11).

### 24.2 Enzymology of RNA Synthesis: RNA Polymerase

As discussed earlier, RNA synthesis involves copying a template DNA strand. However, the earliest known enzyme capable of RNA synthesis in vitro did not require a template. This RNA-synthesizing enzyme, called polynucleotide phosphorylase, was discovered in the 1950s. The enzyme was quite different from DNA polymerase. The enzyme required no template, and it used ribonucleoside diphosphates (rNDPs) as substrates to produce a random-sequence polynucleotide whose base composition matched the nucleotide composition of the reaction medium.

\[
\text{n rNDP} \rightleftharpoons (\text{rNMP})_n + n \text{ P}_i
\]

Initially, it was thought that polynucleotide phosphorylase might be the major RNA-synthesizing enzyme, but the lack of a template requirement was puzzling, as was the apparent absence of the enzyme in eukaryotic cells. Ultimately, polynucleotide phosphorylase turned out to play no role in RNA synthesis in vivo but instead was found to participate in the degradation of bacterial mRNAs. The enzyme was of great value, however, in the synthesis of polynucleotides used as templates for in vitro protein synthesis, when the genetic code was being elucidated (see Chapter 25).

Investigators continued to search for an enzyme that would copy a DNA template in vitro. In 1960, such an enzyme was discovered almost simultaneously in four different laboratories. The enzyme, DNA-directed RNA polymerase, resembled DNA polymerases in the nature of the reaction catalyzed.

\[
\text{n (ATP + GTP + CTP + UTP)} \xrightleftharpoons{\text{Mg}^2+,\text{DNA}} (\text{AMP} - \text{GMP} - \text{CMP} - \text{UMP})_n + n \text{ PP}_i
\]

The reaction product is a complementary RNA copy of the DNA template.

**Biological Role of RNA Polymerase**

In bacteria a single RNA polymerase catalyzes the synthesis of mRNA, rRNA, and tRNA. This was shown in experiments with rifampicin, an antibiotic that inhibits RNA polymerase in vitro and blocks the synthesis of mRNA, rRNA, and tRNA in vivo. Rifampicin-resistant mutants of *E. coli* were found both to contain a rifampicin-resistant form of RNA polymerase and to be capable of synthesizing all three RNA classes in vivo in the presence of rifampicin. Because a single mutation affects both the RNA polymerase and the synthesis of all RNA types in vivo, RNA polymerase must be the one enzyme catalyzing all forms of transcription in bacteria.

**CONNECTION** Rifampicin is a widely used antibiotic that specifically targets bacterial RNA polymerase.

In contrast, eukaryotes contain three distinct RNA polymerases, one each for synthesis of tRNA, mRNA, and small RNAs (tRNA plus the 5S species of rRNA)—RNA polymerases I, II, and III, respectively. The existence of separate enzymes was revealed partly because they differ in their sensitivity to inhibition by α-amanitin, a toxin from the poisonous *Amanita* mushroom. RNA polymerase II is inhibited at low concentrations of α-amanitin, RNA polymerase III is inhibited only at high concentrations, and RNA polymerase I is quite resistant.

Because DNA polymerases and RNA polymerases catalyze similar reactions, it is worthwhile to compare some of their kinetic features.
The $k_{cat}$ for *E. coli* DNA polymerase III holoenzyme, at about 500 to 1000 nucleotides per second, is much higher than $k_{cat}$ for purified RNA polymerase—50 nucleotides per second, which is about the same as the rate of transcription in vivo. Although there are only about 10 molecules of DNA polymerase III per *E. coli* cell, there are some 2000 molecules of RNA polymerase, of which half might be involved in transcription at any instant. Why is this significant? As seen in Chapter 22, replicative DNA chain growth is rapid but takes place at few sites, whereas transcription is much slower but occurs at many sites. The result is that far more RNA accumulates in the cell than DNA. Like the DNA polymerase III holoenzyme, the action of RNA polymerase is highly processive. Once past the initial stages of transcription, RNA polymerase rarely, if ever, dissociates from the template until the specific signal to terminate has been reached.

Another important difference between DNA and RNA polymerases is the accuracy with which a template is copied. With an error rate of about $10^{-15}$, RNA polymerase is less accurate than replicative DNA polymerase holoenzymes, although RNA polymerase is much more accurate than expected if base-pairing with template nucleotides were the sole determinant of the base sequence of the transcript. Given that RNA does not carry information from one cell generation to the next, an ultrahigh-fidelity template-copying mechanism is evidently not needed.

**Structure of RNA Polymerase**

When highly purified *E. coli* RNA polymerase is analyzed in denaturing electrophoretic gels, five distinct polypeptide subunits are observed. Their properties are summarized in Table 24.1. Two copies of the α subunit are present, along with one each of β, β', σ, and ω, giving $M_r$ of about 450,000 for the holoenzyme.

Table 24.1 Subunit composition of *E. coli* RNA polymerase

<table>
<thead>
<tr>
<th>Subunit</th>
<th>$M_r$</th>
<th>Number per Enzyme Molecule</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>α</td>
<td>36.5 kDa</td>
<td>2</td>
<td>Chain initiation, interaction with regulatory proteins and upstream promoter elements</td>
</tr>
<tr>
<td>β</td>
<td>151.0 kDa</td>
<td>1</td>
<td>Chain initiation and elongation</td>
</tr>
<tr>
<td>β'</td>
<td>155.0 kDa</td>
<td>1</td>
<td>DNA binding</td>
</tr>
<tr>
<td>σ</td>
<td>70.0 kDa$^a$</td>
<td>1</td>
<td>Promoter recognition</td>
</tr>
<tr>
<td>ω</td>
<td>11.0 kDa</td>
<td>1</td>
<td>Promotion of enzyme assembly</td>
</tr>
</tbody>
</table>

$^a$ The 70-kDa σ subunit is one of several alternative σ subunits.

The σ subunit is easily dissociated from RNA polymerase. The σ-free enzyme, called core polymerase, is still catalytically active, but it binds to DNA at far more sites than does the RNA polymerase holoenzyme, and it shows no strand or sequence specificity. The σ subunit thus plays an essential role in directing RNA polymerase to bind to the duplex DNA at the proper site for initiation—the promoter site—and to select the correct strand as the template for transcription. The addition of σ to core polymerase reduces the affinity of the enzyme for nonpromoter sites by about 10$^7$, thereby increasing the enzyme's specificity for binding to promoters.

These discoveries about σ suggested that gene expression could be regulated by having core polymerase interact with different forms of σ, which would in turn direct the holoenzyme to different promoters. In fact, this does occur. In one example an *E. coli* culture is stressed by a sudden temperature increase. In these *heat-shocked* cells, a new form of σ appears and directs the modified RNA polymerase to a different set of promoters, thereby activating transcription of a block of genes called heat-shock genes. The most abundant σ in *E. coli*, and the one that will frame our discussions, is called σ$^{70}$ because of its 70-kDa molecular weight. Subunit σ$^{70}$ is one of seven different σ factors known in *E. coli*, each designed to direct RNA polymerase to a functionally related set of genes.

Eukaryotic RNA polymerases show similarity with the bacterial enzymes, but they have a more complex subunit structure. RNA polymerase II (pol II) from yeast has 12 subunits. The common ancestry among RNA polymerases is evident, however, as shown in Figure 24.4.

**Figure 24.4**

RNA polymerase subunit structures in the three domains of life. The subunits are arranged by function rather than by size. Homologous subunits are color-coded. Subunits marked with an asterisk are conserved among the three eukaryotic RNA polymerases. Only core enzyme subunits are shown. Data from F. Werner (2007) Structure and function of archaean RNA polymerases, Molecular Microbiology 65:1395–1404.
For example, the two largest subunits in the three eukaryotic multisubunit polymerases are related to $\beta$ and $\beta'$ of the bacterial enzyme. The figure also shows the subunit composition of archaeal RNA polymerases, which reveal them to be much more closely related to eukaryotic than to bacterial polymerases. Eukaryotic RNA polymerases have no direct counterpart to bacterial $\sigma$. Instead, a series of proteins called transcription factors functions in comparable fashion, helping to direct RNA polymerase to promoter sites, forming an initiation complex (page 713).

Although the multisubunit motif for RNA polymerases is the dominant structural theme, it is not universal. Exceptions include single-subunit polymerases encoded by some bacteriophage genomes, including T7 and SP6, and a mitochondrial RNA polymerase found in vertebrates. Plant chloroplasts also contain an organelle-specific RNA polymerase as well as two additional RNA polymerases, IV and V. These enzymes are involved in RNA-directed DNA methylation, a process in gene silencing (Chapter 26).

The crystal structures of eukaryotic and bacterial RNA polymerases were first determined in 2001 and 2002 in the laboratories of Roger Kornberg and Seth Darst, respectively. **Figure 24.5(a)** shows the structure of *E. coli* RNA polymerase holoenzyme (containing sigma), reported in 2013, and **Figure 24.5(b)** shows a cutaway view of yeast RNA polymerase II, illustrating structural features common to all of the multisubunit RNA polymerases. We shall describe the relationship between these structures and the RNA polymerase reaction presently.

### 24.3 Mechanism of Transcription in Bacteria

Like DNA replication and protein synthesis, transcription occurs in three distinct phases—initiation, elongation, and termination. Initiation and termination signals in the DNA sequence punctuate the genetic message by directing RNA polymerase to particular genes and by specifying where transcription will start, where it will stop, and which DNA strand will be transcribed. The signals involve both instructions encoded in DNA base sequences and interactions between DNA and proteins other than RNA polymerase. We focus first on bacterial RNA polymerases, exemplified by the widely studied *E. coli* enzyme, but the basic mechanics of transcription are similar in all organisms. Actions of the eukaryotic enzymes are discussed later.

#### Initiation of Transcription: Interactions with Promoters

Promoters, the DNA sites at which RNA polymerase binds and initiates transcription, were originally identified from DNA base sequence analysis upstream (on the 5′ side) from sites at which transcription begins. This analysis for several *E. coli* genes identified two conserved adenine- and thymine-rich sequences centered at about 10 and 35 nucleotides to the 5′ side of the transcriptional start site, or, at positions −10 and −35, where +1 represents the first DNA nucleotide to be

![Crystal structure of RNA polymerase.](image)

**FIGURE 24.5**
transcribed (FIGURE 24.6). There was some sequence variation among the promoters analyzed, but a consensus sequence emerged—meaning those nucleotides found most often at each position in each site—within this conserved region.

Biochemical analysis by a technique called footprinting confirmed that RNA polymerase binds tightly to DNA sequences immediately upstream from the transcription start site and that the -10 and -35 “boxes” are in close contact with the bound enzyme. In footprinting, a cloned segment of DNA is made radioactive by use of a kinase enzyme and \[^{32P}\text{ATP}\] to phosphorylate the DNA molecule at its 5′ end. The 5′ end-labeled DNA is mixed with RNA polymerase or other site-specific DNA-binding protein of interest. As shown in FIGURE 24.7, the end-labeled DNA by itself and the DNA–protein complex are digested separately with pancreatic deoxyribonuclease (DNase I), an enzyme that cleaves DNA at random sites. Conditions are chosen so that most DNA molecules are cleaved once. The reaction mixtures are then subjected to gel electrophoresis and autoradiography. DNA sites in close contact with the protein are protected from cleavage, and bands representing those sites are not seen.

The search process by which RNA polymerase finds a promoter site embedded in many kilobases of DNA is not well understood. or is required, for core polymerase binds weakly to DNA with no specificity whatsoever. The initial encounter between RNA polymerase holoenzyme and a promoter generates a closed-promoter complex. Whereas DNA strands unwind later in transcription, no unwinding is detectable in a closed-promoter complex. Binding is primarily electrostatic and of low affinity. Footprinting studies show that polymerase is in contact with DNA from about nucleotide -55 to -5.

Next, RNA polymerase unwinds several base pairs of DNA, from about -12 to -1, giving an open-promoter complex, so-called because it binds DNA whose strands are open, or

### FIGURE 24.6
Conserved sequences in promoters recognized by *E. coli* RNA polymerase. Lengths of spacer regions between the two conserved segments are shown.

<table>
<thead>
<tr>
<th>Promoter for:</th>
<th>-35 region</th>
<th>Spacer</th>
<th>-10 region</th>
<th>Spacer</th>
<th>Transcribed</th>
</tr>
</thead>
<tbody>
<tr>
<td>trp operon</td>
<td>G T T G A C A</td>
<td>N\text{17}</td>
<td>T T A A C T</td>
<td>N\text{2}</td>
<td>A</td>
</tr>
<tr>
<td>tRNA\text{(Tyr)} operon</td>
<td>C T T T A C A</td>
<td>N\text{16}</td>
<td>T A T G A T</td>
<td>N\text{2}</td>
<td>A</td>
</tr>
<tr>
<td>(\lambda)P2</td>
<td>G T T G A C A</td>
<td>N\text{17}</td>
<td>G A T A C T</td>
<td>N\text{2}</td>
<td>G</td>
</tr>
<tr>
<td>lac operon</td>
<td>C T T T A C A</td>
<td>N\text{17}</td>
<td>T A T G T T</td>
<td>N\text{2}</td>
<td>A</td>
</tr>
<tr>
<td>rec(\lambda)</td>
<td>C T T G A T A</td>
<td>N\text{16}</td>
<td>T A T A A T</td>
<td>N\text{2}</td>
<td>A</td>
</tr>
<tr>
<td>lex(\lambda)</td>
<td>G T T C C A A</td>
<td>N\text{17}</td>
<td>T A T A C T</td>
<td>N\text{2}</td>
<td>A</td>
</tr>
<tr>
<td>T7A3</td>
<td>G T T G A C A</td>
<td>N\text{17}</td>
<td>T A C G A T</td>
<td>N\text{2}</td>
<td>A</td>
</tr>
</tbody>
</table>

**CONCEPT** Transcription begins with sequence-specific interaction between RNA polymerase and a promoter site, where duplex unwinding and template strand selection occur.
unwound. The open-promoter complex is extremely stable; it forms with a $K_a$ as high as $10^{12}$ M$^{-1}$. Structural analysis of DNA complexed with $\sigma$ shows specific interactions between the protein and DNA bases in the $-10$ box. DNA melting is an essential feature of promoter recognition.

Next, an Mg$^{2+}$-dependent isomerization occurs, giving a modified form of the open-promoter complex with the unwound DNA region now extending from $-12$ to $+2$. DNA is bent in this structure, as shown in Figure 24.8. Note also the positions of the major promoter elements ($-10$ and $-35$) and the bound Mg$^{2+}$ at the catalytic site.

### Initiation and Elongation: Incorporation of Ribonucleotides

RNA chain growth begins with binding of the template-specified rNTP, followed by binding of the second nucleotide and its fitting to the template. Nucleophilic attack by the 3′ hydroxyl of the first nucleotide on the α (innermost) phosphorus of the second nucleotide generates the first phosphodiester bond and leaves an intact triphosphate moiety at the 5′ position of the first nucleotide. Nucleotide incorporation occurs via the two-metal mechanism described for DNA polymerases in Chapter 22. The transcript is bound unstably during the first several phosphodiester bond-forming reactions, as shown by the fact that most initiations are abortive, with release of oligonucleotides two to nine residues long. The basis for this low efficiency of initiation is not fully understood.

During incorporation of the first 10 nucleotides, the $\sigma$ subunit dissociates from the transcription complex, and the remainder of the transcription process is catalyzed by the core polymerase. Once $\sigma$ has dissociated, the elongation complex becomes quite stable. Transcription, as studied in vitro, can no longer be inhibited by adding rifampicin, and virtually all transcription events proceed to completion.

During elongation, the core enzyme moves along the duplex DNA template. As it moves, it simultaneously unwinds the DNA, exposing a single-stranded template for base pairing with incoming nucleotides and with the nascent transcript (the most recently synthesized RNA); it rewinds the template behind the 3′ end of the growing RNA chain, as suggested in Figure 24.8. In this model, about 18 base pairs of DNA are unwound at any given time, forming a moving “transcription bubble.” As one base pair becomes unwound in advance of the 3′ end of the nascent RNA strand, one base pair becomes rewound near the trailing end of the RNA polymerase molecule. About nine base pairs of the 3′ end of the nascent transcript are hybridized to the template DNA strand. During elongation, RNA polymerase functions as a true molecular motor. Techniques for analyzing single complexes show that RNA polymerase generates forces exceeding those of well-studied cytoskeletal motor proteins, such as myosin and kinesin.

Refer back to Figure 24.5(b) for a stylized picture of an elongation complex. Although this model was derived from the structure of yeast RNA polymerase II, it shows generally applicable features of transcriptional elongation, including the large “wall,” which forces DNA to bend, almost at right angles. In this schematic illustration, the direction of polymerase motion is left to right as shown. DNA entering the enzyme is gripped by protein “jaws” (upper jaw not shown in this cutaway model). The 3′ end of the growing RNA is adjacent to one of the catalytically essential Mg$^{2+}$ ions. The wall forces the DNA to turn. rNTPs probably enter the active site, as shown, through a funnel structure and pore. The 5′ end of the growing RNA chain is diverted from the DNA template by a protein loop called the rudder, which limits the length of RNA hybridized to template DNA. The rudder and lid, which guide the exit of RNA, emanate from a large clamp that swings from back to front, as shown, over the catalytic site and contributes to binding nucleic acids, and hence, to the high processivity of transcription.

The concept of a transcription bubble as a central intermediate in transcription, as depicted in Figure 24.9, suggests that the enzyme moves along the DNA template in register with the growing RNA transcript, with the footprint advancing by one base pair for each ribonucleotide incorporated into the transcript. However, footprinting of numerous initiation and elongation complexes has shown that the enzyme often advances discontinuously, holding its position for several cycles of nucleotide addition and then jumping forward by several base pairs along the template. These and other observations suggest that the means of RNA polymerase translocation is fundamentally different from the continuous movement implied by a picture of the transcription bubble. It has long been known that some DNA sequences are difﬁcult to transcribe, and RNA polymerase “pauses” when it reaches...
such a site in vitro, often sitting at the same site for several seconds before transcription is resumed. At such sites, RNA polymerase often translocates backward, and in the process the 3’ end of the nascent transcript is displaced from the catalytic site of the enzyme. This gives a 3’ “tail,” which may be several nucleotides long and is not base-paired to the template, protruding downstream of the enzyme (FIGURE 24.10). For transcription to resume, an RNA 3’ end must be positioned in the active site. At this point the enzyme may either cleave the single-stranded RNA end or move forward without new nucleotide incorporation, thereby repositioning the RNA 3’ nucleotide in duplex conformation. These pausing, backtracking, and cleavage reactions are thought to help control the transcription rate and to enhance transcriptional fidelity by removing mismatched or damaged ribonucleotides.

**Punctuation of Transcription: Termination**

Because of the great stability of transcription complexes, termination of transcription, with release of the nascent transcript, is an involved process. In bacteria we recognize two distinct types of termination events—those that depend on the action of a protein **termination factor** called $\rho$ (rho), and those that are $\rho$ factor–independent.

**Factor-Independent Termination**

Sequencing the 3’ ends of genes that terminate in a factor-independent manner reveals two structural features shared by many such genes and is illustrated in FIGURE 24.11: (1) two symmetrical GC-rich segments

- ... CCCA GCC GCC CUA UCA G...
- 3’-GGGT CGGG CGGA TAC T...CTT TTT-5’
- 5’- CCCA GCC GCC CUA UCA G...
- GCGGCC UUU UUU-UH$_3^*$

(a) An A-rich segment of the template (orange segment on right) has just been transcribed into a U-rich mRNA segment.

Complementary GC-rich parts of the transcript base-pair with one another, displacing this part of the transcript from its template or from its enzyme-binding site.

- 3’-GGGT CGGG CGGA TAC T...CTT TTT-5’
- 5’- CCCA GCC GCC CUA UCA G...
- GCGGCC UUU UUU-UH$_3^*$

(b) The RNA–RNA duplex, stabilized by G–C base pairs (yellow), eliminates some of the base pairing between template and transcript.

- 3’-GGGT CGGG CGGA TAC T...CTT TTT-5’
- 5’- CCCA GCC GCC CUA UCA G...
- GCGGCC UUU UUU-UH$_3^*$

(c) The unstable A–U bonds linking transcript to template hybrid dissociate, releasing the transcript.

**FIGURE 24.10**

Backtracking in an elongation complex. Above, the 3’ terminus of the transcript is in the catalytic site. Below, the enzyme has slipped backward, leaving the 3’ transcript terminus at the end of a non–base-paired RNA tail, several nucleotides long. Data from E. Nudler (2012). RNA polymerase backtracking in gene regulation and genome instability. Cell 149:1438–1445.

**FIGURE 24.11**

A model for factor-independent termination of transcription.
RNA polymerase

mRNA

Binding of ρ to DNA–RNA–polymerase complex

5' mRNA

ρ protein

3'

RNA polymerase

ρ recognition site

ρ moves toward 3’ end, displacing the DNA template strand

5' ATP

ADP + P_i

3'

This weakens the interaction between template and transcript, causing them to dissociate; ρ and polymerase also dissociate

CONCEPT DNA sequences that promote factor-independent termination include a GC-rich region that forms a stem-loop followed by a run of 4 to 8 A residues.

CONCEPT In factor-dependent termination, ρ protein acts as an RNA–DNA helicase, unwinding the template–transcript duplex and facilitating release of the transcript.

24.4 Transcription in Eukaryotic Cells

Transcription in eukaryotes is more complex than in bacteria and their phages. Not only is there much more discrimination in what is to be transcribed and what is not, but this transcription must be precisely programmed during development and tissue differentiation. Furthermore, the transcription machinery must deal with the complex levels of structure in eukaryotic chromatin. Reflecting this complexity is the fact that eukaryotic cells have three different RNA polymerases, each with a specialized function (we leave aside the two additional polymerases in plants). For each polymerase several proteins must assemble at promoters and other upstream sites on the template DNA, along with RNA polymerase, in order to form a functional transcription complex. None of the three polymerases has a direct counterpart to the σ factor of bacterial complexes. However, all three require a set of transcription factors that play
CONCEPT Eukaryotic cells have three RNA polymerases, each requiring additional protein factors to initiate transcription. Transcription factors are named TFI, TFII, or TFIII, depending on whether they function with RNA polymerase I, II, or III, respectively. Within one class of transcription factors, each individual factor is identified with a letter; thus, TFI1A is one of several transcription factors functioning with RNA polymerase II. The ensemble of proteins required to form an initiation complex is conserved among polymerases I, II, and III.

Other differences from bacteria pertain to the fact that bacterial genomes are organized into blocks of functionally related genes—operons, such as the lactose operon mentioned earlier in this chapter—which are co-transcribed to give multicistronic mRNAs, while eukaryotic genes are almost always transcribed singly and processed as templates for individual proteins. Post-transcriptional processing, as discussed later in this chapter, is far more complex for eukaryotic than for bacterial transcripts.

RNA Polymerase I: Transcription of the Major Ribosomal RNA Genes

The eukaryotic ribosome contains four rRNA molecules (see Chapter 25). The small subunit has an 18S rRNA, whereas the large subunit contains 28S, 5.8S, and 5S components. Of these, the 28S, 18S, and 5.8S subunits are all produced from an initial 45S pre-rRNA transcript, and it is the special function of RNA polymerase I (pol I) to carry out this transcription. rRNA transcription and processing, along with ribosome assembly, occur in the nucleolus. Transcription occurs from multiple, tandemly arranged copies of the 45S rRNA gene, as shown in FIGURE 24.13. After transcription, the 45S pre-rRNA is processed to yield 18S, 5.8S, and 28S rRNA molecules. About 6800 nucleotides are discarded in this process. The rRNAs are then combined with 55 RNA from the nucleus and ribosomal proteins synthesized in the endoplasmic reticulum. The resulting ribosomal subunits are exported from the nucleolus back into the cytosol.

CONCEPT Pol I transcribes the major ribosomal RNA genes; pol III transcribes small RNA genes; and pol II transcribes protein-encoding genes and a few small RNA genes.

RNA Polymerase III: Transcription of Small RNA Genes

RNA polymerase III (pol III) is the largest and most complex of the eukaryotic RNA polymerases (see Figure 24.4). The major targets for pol III are the genes for all the tRNAs and for the ribosomal 5S rRNA. Like the major ribosomal genes described in the previous section, these small genes are present in multiple copies, but they are usually not grouped together in tandem arrays, nor are they localized in one region of the nucleus. Rather, they are scattered over the genome and throughout the nucleus.

Of all the genes transcribed by pol III, the most thoroughly studied are those for 5S rRNA. At least three transcription factors—TFIIB, –B, and –C, are needed for the expression of these genes. TFIIB and TFIIC participate in transcribing tRNA genes as well, but TFIIB is specific only for the 5S genes. TFIIB is an example of an abundant class of sequence-specific DNA-binding proteins, in which metal-binding zinc fingers make contact with and identify DNA sequences (FIGURE 24.14). This class of proteins contains conserved histidine and cysteine residues, which complex with Zn²⁺. This DNA-binding protein motif was mentioned in Chapter 20 as a structural element in steroid hormone receptors.

RNA Polymerase II: Transcription of Structural Genes

All of the structural genes (those encoding proteins) in the eukaryotic cell are transcribed by RNA polymerase II (pol II). This enzyme also transcribes some of the small nuclear RNAs involved in splicing (discussed in Section 24.5). Like other RNA polymerases, pol II is a complex multisubunit enzyme. However, not even its 12 subunits are sufficient to allow pol II to initiate transcription on an eukaryotic promoter. Because the expression of many eukaryotic genes is either tissue-specific or developmental stage-specific or both, eukaryotic promoter structure is far more complex than that of bacteria. Protein factors in addition to RNA polymerase are required for promoter recognition, recruitment of RNA polymerase to a promoter, and generation of an active elongation complex. A typical initiation complex contains 60 proteins in addition to RNA polymerase. Of these, about half form a preinitiation complex that assists the enzyme in recognizing and binding to a promoter. The remainder are involved in regulation.

A typical eukaryotic promoter contains an initiator region (Inr) with the sequence YYANWYY, where N is any nucleotide, Y is a pyrimidine (C or T), W is either A or T, and N represents the +1 initiation site.
A counterpart to the bacterial –10 region, called the TATA box, positioned between –20 and –30, has the sequence TATAAAA. Upstream from that box are arrayed additional control elements, including the CAAT box (GGCCAATCT), the GC box (GGGCGG), and Octamer (ATTTGCAT). Additional regulatory sites may exist several kbp upstream from the initiation site; these are called enhancer regions. Although these far upstream activator sites are involved in transcriptional regulation, they are not considered part of the promoter itself. **Figure 24.15** shows the locations of these elements in several well-studied eukaryotic promoters.

RNA polymerase II interacts with several general transcription factors, including TBP (TATA box-binding protein) and TFIIA, -B, -E, -F, and -H. The formation of initiation and elongation complexes has been studied in detail with yeast RNA polymerase II, by use of crystallographic and other biophysical methods. A remarkable feature of the preinitiation complex, as shown recently by Roger Kornberg’s laboratory, is the fact that DNA is bound in this complex but is not in direct contact with RNA polymerase.

TFIIB plays an important role in converting the initial closed-promoter complex to an open-promoter complex. The process, summarized below, is illustrated in a movie produced in Patrick Cramer’s laboratory. (See reference to Cheung and Cramer at the end of this chapter.) **Figure 24.16** shows scenes from the movie. Panel (a) shows a front view of the core RNA polymerase, while the other panels present side views at different stages of the process.

In forming the closed-promoter complex, TBP first binds to DNA and bends it by 90 degrees (Figure 24.16(b)). The C-terminal domain of TFIIB binds to TBP and flanking DNA regions. The N-terminal domain recruits RNA polymerase to promoter DNA near the transcription start site, forming the closed-promoter complex, also shown in panel (b). Next, a TFIIB element called the B-linker opens DNA before the transcription start site, leading to an open-promoter complex (Figure 24.16(c)). ATP is required for this process. TFIIB threads the template DNA strand into the active center. For this TFIIB uses another structural element, the B-reader, which consists of a helix followed by a mobile loop. Next, DNA is scanned for an initiator (Inr) motif near the start site; DNA “scrunching” movements of the non-template strand facilitate this process. Following this, the first two ribonucleotide substrates are positioned opposite Inr, and the first phosphodiester bond is formed (Figure 24.16(d)). Just as seen with bacterial RNA polymerase, most of the early chain initiation events are abortive. Finally, growth of RNA chains beyond seven nucleotides triggers release of TFIIB, and this completes the process of promoter escape. The process as described is actually similar to transcription initiation as studied with bacterial RNA polymerases, even though the proteins other than RNA polymerase are quite different.

**Figure 24.14** Zinc fingers. Data from M. S. Lee et al (1989) Science 245:635–637.

**Figure 24.15** Structures of some eukaryotic promoters. The colored boxes represent different regulatory elements: orange, TATA box; blue, GC box; yellow, CAAT box; purple, Octamer. Based in part upon Genes IV, B. Levin, Oxford: Oxford University Press, 1990.
As mentioned earlier, trans-acting factors binding at enhancer sequences far removed from the promoter itself—by as much as several kilobase pairs—can influence transcription. Their mode of action appears to involve DNA looping, perhaps mediated by nucleosomes, which can bring enhancer-bound proteins into close physical contact with proteins bound to the promoter. Some transcription factors can bind in either promoter or enhancer regions. These can act as intermediates between activators or repressors bound to enhancer regions and the core transcription complex as schematized in Figure 24.17. Also involved in communication between upstream control elements and proteins bound at the promoter is a multiprotein complex called mediator (Chapter 26).
Transcription in Eukaryotic Cells

Figure 24.17

A schematic representation of DNA looping as a process to bring enhancer-bound (ENH) activator (Act) proteins into contact with trans-acting factors (TAFs) associated with the core transcription complex.

Chromatin Structure and Transcription

The complex interplay of transcription factors and polymerases we have described occurs not on naked DNA but on chromatin. The chromatin structure presents two major problems: First, how can the transcription factors and initiation complex bind to DNA in the presence of nucleosomes? Second, how can the actively transcribing polymerase pass through arrays of nucleosomes? This is an area of intense research interest, which we treat in more detail in Chapter 26. Here we can point out that transcription often initiates in chromatin regions containing nuclease-accessible sites—regions in which the DNA in isolated chromatin is readily cleaved, as if it were not complexed with histones at those sites, granting ready access to RNA polymerase.

How are accessible sites established in previously unresponsive genes? In some cases, such as the globin genes, it seems that the chromatin structure is rearranged at the time of replication. In other instances, protein factors seem able to interfere with chromatin structure at specific loci, opening hypersensitive sites. In either case, the clearance of histones from nuclease-susceptible sites involves the action of chromatin remodeling factors. These are proteins that enable promoter regions to be able to accept the complex and bulky machinery depicted in Figure 24.16. These complexes, which we discuss in Chapter 26, require ATP hydrolysis to somehow “open” nucleosomes transiently, to allow transcription complexes to form.

Another and perhaps equally important role is played by histone acetyltransferases and deacetylases. Histones of the nucleosome core are subject to acetylation at specific lysine residues in the N-terminal tails (see Figure 24.18); chromatin is subject to other modifications as well, as mentioned in Chapter 21. High levels of acetylation are correlated with high transcriptional activity, and low acetylation with low activity. Chemically, this makes sense: neutralization of histone basic residues by acetylation would loosen ionic interactions between histones and DNA in chromatin. A number of proteins recruited to the initiation complex by activators and trans-acting factors have histone acetylase activity. The fact that specific transcription factors are involved in this process may provide the long-sought explanation for how the chromatin of specific genes can be targeted for disruption.

Transcriptional Elongation

Formation of the open-promoter complex (see Figure 24.16) is followed, in the presence of rNTPs and ATP, by melting of a short region of DNA and initiation of transcription. As noted previously, the
C-terminal tail of the Rpb1 subunit of pol II becomes highly phosphorylated, leading to promoter release, and elongation begins, with a helicase activity clearing the way. A number of the core transcription factors are released, and pol II, together with TFIIF, moves along the DNA. A residual complex, containing TBP, TFIIB, TAFs, and probably activator proteins, remains at the start site, ready to initiate another round.

At this point the polymerase also acquires several elongation factors. Some of these factors assist the enzyme in traversing pause sites in the DNA. As previously noted for bacterial RNA polymerase, transcription is relatively slow and interrupted by frequent pauses, especially in T-rich regions. Elongation factors assist the enzyme in passing such sites. Nucleosomes form even larger obstacles to the progress of an RNA polymerase II along the DNA, as mentioned previously.

Just how pol II transcribes through nucleosomes is still something of a mystery. Do the nucleosomes unfold and re-form as the polymerase passes? Are they temporarily displaced? Current evidence favors temporary displacement, but the issue is far from settled.

Termination of Transcription
The termination of mRNA transcription is also different in eukaryotes. Whereas the bacterial RNA polymerase recognizes terminator signals, which sometimes function with the aid of the ρ protein, the eukaryotic polymerase II usually continues to transcribe well past the end of the gene. In doing so, it passes through one or more TTATTT signals, which lie beyond the 3′ end of the coding region (see FIGURE 24.19). The pre-mRNA, carrying this signal as AAUAAA, is then cleaved by a special endonuclease that recognizes the signal and cuts at a site 11 to 30 residues downstream of it. At this point, a tail of polyriboadenyllic acid, poly(A), as many as 300 bases long, is added by sequential ATP incorporation, catalyzed by a special nontemplate-directed enzyme, poly(A) polymerase. The functions of the poly(A) tails of eukaryotic mRNAs include mRNA stabilization and facilitation of transport from nucleus to cytoplasm. We know that they cannot be essential for all messages, because some mRNAs (e.g., most histone mRNAs in higher eukaryotes) do not have them. However, poly(A) tails relate to message stability, for tail-less messages typically have much shorter lifetimes in the nucleus. Recent evidence indicates that the specific TTATTT signals used for termination vary in different tissues, so that the 3′ end of an mRNA is partly tissue-specific. The functional significance of this variation is not yet known.

24.5 Post-transcriptional Processing
Bacterial mRNA Turnover
Critical to mRNA metabolism in eukaryotes are events occurring after transcription—events that are necessary for messages to move from the nucleus to their sites of utilization in the cytosol. We discuss these events later in this chapter. In bacteria, by contrast, mRNAs are available for use in protein synthesis immediately. In fact, a nascent mRNA can serve as a template for translation at its 5′ end while still in the process of being synthesized toward the 3′ end. That is, transcription is coupled directly to translation.

The major post-transcriptional event in metabolism of bacterial mRNA is its own degradation, which in most cases is rapid. A few bacterial mRNAs, notably those encoding outer membrane proteins, are long-lived; however, many bacterial messages have half-lives of only 2 to 3 minutes. This short life span means that genes being expressed must be transcribed continuously and that many mRNA molecules are translated only a few times. Although this might seem wasteful, it is consistent with bacterial lifestyles, which necessitate rapid adaptation to environmental changes. Earlier we noted the selective advantage to bacteria of expressing the genes for lactose utilization only if an inducer is present. By the same token, it would be wasteful for the cell to continue producing these proteins after lactose or a related sugar was exhausted from the milieu. Rapid degradation of lac mRNA ensures that the seemingly wasteful synthesis of these proteins will cease soon after the need for these proteins is gone.

Turnover of bacterial ribosomes occurs largely in the degradosome, a multiprotein complex containing polynucleotide phosphorylase (see Section 24.2), a nuclease called RNase E, and an RNA helicase, which helps to unwind regions of mRNA secondary structure. Degradation starts from the 5′ end, which is important because translation also starts from the 5′ end. If degradation were to start from the 3′ end, a ribosome starting from a 5′ end might never reach an intact 3′ end.

Post-transcriptional Processing in the Synthesis of Bacterial rRNAs and tRNAs
Both ribosomal RNAs and transfer RNAs are synthesized as larger transcripts (pre-rRNA and pre-tRNA, respectively), which undergo cleavage at both ends of the transcript, en route to becoming mature RNAs. This process is comparable to the processing of pre-rRNA in eukaryotic cells. As we will see in Chapter 25, however, the RNA components in bacteria are somewhat smaller than in eukaryotes—23S, 16S, and 5S. The total amount of DNA encoding these RNAs
Aside from the tRNAs embedded in pre-rRNA transcripts, the other tRNAs are synthesized in transcripts that contain one to seven tRNAs each, all surrounded by lengthy flanking sequences. The maturation steps are summarized in Figure 24.21, using as an example the well-studied case of the E. coli tyrosine tRNA species (tRNA\textsubscript{Tyr}). Maturation (step 1) starts with RNase E, which cleaves next to a stem-loop structure on the 3' side of the tRNA sequence. This is followed by the action of ribonuclease D (step 2), which carries out exonucleolytic cleavage to a point two nucleotides removed from the CCA sequence at the 3' end. Next, the 5' end is formed by the action of ribonuclease P (step 3), which cleaves to leave a phosphate on the 5' terminal G. This enzyme creates the 5' terminus of all tRNA molecules. It is unclear what structural features are recognized by RNase P, because different sequences are contained in the cleavage sites. As pointed out in Chapter 8, ribonuclease P was one of the first identified ribozymes. The enzyme consists of one RNA molecule of 377 nucleotides and one protein molecule with Mr of about 20,000. Both components are necessary for full catalytic activity, but the RNA molecule alone can catalyze accurate cleavage.

**CONCEPT** Bacterial transcripts undergo post-transcriptional processing, involving both endonucleolytic and exonucleolytic cleavage.
Once the proper 5′ terminus has been formed, ribonuclease D removes the remaining two nucleotides from the 3′ end (step 4). Should excessive “nibbling” occur through faulty control of RNase D activity, there is an enzyme (CCA nucleotidyltransferase) that will restore the CCA end to any tRNA in a nontemplated fashion. This enzyme specifically recognizes the 3′ terminus of tRNAs that lack the CCA end and catalyzes sequential reactions with a CTP, another CTP, and an ATP. Note that the CCA end is encoded by every tRNA gene, so the nucleotidyltransferase is essentially a repair enzyme.

Creation of the modified bases (see Chapters 4 and 27) occurs at the final stage, including methyllations, thiolations, and reduction of uracil to dihydrouracil. In the specific example shown, the modifications include formation of two pseudouridines, one 2-isopentenyladenosine, one O2′-methylguanosine, and one 4-thiouridine (step 5). These modifications serve to stabilize the tRNA molecules against intracellular degradation, and in some cases they promote translational fidelity. The modifications are not essential for tRNA function, however, because many tRNAs lacking the modifications are fully active in vitro. Pathways for eukaryotic tRNA synthesis are similar, including the involvement of ribonuclease P. In the yeast Saccharomyces cerevisiae the average tRNA molecule has more than 12 of its 75–80 bases modified.

### Processing of Eukaryotic mRNA

Bacterial and eukaryotic cells differ significantly in the ways that mRNAs for protein-coding genes are produced and processed. Recall that bacterial mRNAs are synthesized at the nucleoid in direct contact with the cytosol and are immediately available for translation. A specific nucleotide sequence at the 5′ end recognizes a site on the bacterial rRNA, allowing attachment of the ribosome and initiation of translation, often even before transcription of the message is completed. Hence, there is little or no post-transcriptional processing of bacterial mRNAs.

In eukaryotes, mRNA is produced in the nucleus and must be exported to the cytosol for translation. Furthermore, the initial product of transcription (pre-mRNA) includes all of the introns and substantial flanking regions; the introns must be removed before correct translation can occur. For these reasons, eukaryotic mRNA requires extensive processing before it can be used as a template. This processing, including polyadenylation at the 3′ end (Section 24.4), takes place while mRNA is still in the nucleus.

### Capping

The first modification occurs at the 5′ end of the pre-mRNA. First, one phosphate is removed hydrolytically from the triphosphate moiety at the 5′ terminal nucleotide. Next, the resulting 5′ diphosphate end attacks the α (inner) phosphate of GTP; in essence, the guanine nucleotide is added in reverse orientation (5′ → 5′). Together with the first two nucleotides of the chain, it forms a cap (FIGURE 24.22). The cap is further modified by the addition of methyl groups to the N-7 position of the guanine and to one or two sugar hydroxyl groups of the cap nucleotides. This cap structure positions the mRNA on the ribosome for translation, and also it probably contributes to stabilization of the message.

---

**CONCEPT** Post-transcriptional processing in bacteria involves cleavage of the primary transcript, modification of bases (in tRNA synthesis) and non-transcriptive nucleotide addition.

**Splicing**

After being capped, the pre-mRNA becomes complexed with a number of small nuclear ribonucleoprotein particles (snRNPs, often called “snurps”), which are themselves complexes of small nuclear RNAs (snRNAs) and special splicing proteins. The snRNAs are all less than 300 nucleotides long. The snRNP–pre-mRNA complex is called a spliceosome, and it is here that the most elegant part of the processing takes place—the cutting and splicing that is necessary to excise introns from the pre-mRNA and join the ends of the two exons. In forming a spliceosome, snRNAs recognize and bind intron–exon splice sites by means of complementary sequences (FIGURE 24.23). Precise recognition
of splice sequences is essential, because even a single-base error would disrupt the sense of the genetic message. A schematic view of the chemistry of splicing is shown in **Figure 24.24**.

Excision of a single intron involves assembling and disassembling a spliceosome. **Figure 24.25** depicts the overall process. The sequence begins with binding of the U1 snRNP to the G site at the 5′ end of the intron. The U2 snRNP then binds at the branch site. With continued assembly of the spliceosome, including the addition of several more snRNPs, the lariat loop in the intron is formed and the two exons are joined. Catalytically essential Mg^{2+} ions are located in the RNA of U6 snRNP, implying that RNA plays a catalytic role. Splicing has now been accomplished, and the products—a ligated mRNA and a looped intron—are released. As the spliceosome disintegrates, the looped intron is degraded, and the mRNA is exported from the nucleus.

Early steps in spliceosome assembly require more proteins and sites than are indicated in Figure 24.25. Kinetic experiments indicate that all of the early reactions in spliceosome assembly are reversible, with the process being driven by the irreversibility of the late steps.

Because of the importance of accurate splicing for correct expression of genetic information, splicing errors are responsible for many genetic diseases. As many as 15% of all genetic diseases arise from splicing errors. In some forms of thalassemia, a family of diseases arising from defective synthesis of hemoglobin chains (Chapter 7), mutations have been found in both the 5′ and 3′ splice sites of both genes for the β chain of human hemoglobin. Usually, an incorrect mRNA chain is formed, which leads to premature termination of translation of the message. Another example of defective splicing is Hutchinson–Gilford progeria syndrome, a disease of accelerated aging in which afflicted individuals...
die in their teens from aging-related symptoms such as cardiovascular disease. The disease affects a nuclear envelope protein called Lamin A. The mutation activates a cryptic donor splicing site in the gene for a lamin A precursor, leading to accumulation of a truncated form of the protein.

**Alternative Splicing**

Once investigators discovered and described mRNA splicing, they were surprised to learn that the same pre-mRNA can undergo splicing in several different ways. The existence of alternative splicing means that different combinations of exons from the same gene can be processed into different mature mRNAs and then undergo translation into quite different proteins in different tissues or at different developmental stages of the same organism. Alternative splicing made it less surprising when the Human Genome Project revealed the existence of far fewer genes than had been expected given the size of the genome and the complexity of *Homo sapiens*. Alternative splicing greatly enlarges the repertoire of proteins that can be encoded by a genome.

A dramatic example of alternative splicing is shown in Figure 24.26. The protein α-tropomyosin is used in contractile systems in various cell types. Apparently, the need for functional domains coded for by different exons differs among the various uses of α-tropomyosin. Rather than having different genes expressed in different tissues, a single gene is employed, but the specific splicing patterns in different tissues provide a variety of α-tropomyosins. As the figure shows, there are two positions at which alternative choices can be made for which exon to splice in. The 3′ member of each of these pairs is the default exon; it will be chosen unless a specific cellular signal dictates otherwise.

Many mechanisms have been described for alternative splicing. A well-studied example involves calcitonin, a peptide hormone that in thyroid gland regulates calcium and phosphorus metabolism. In neuronal tissue, the gene for calcitonin is alternatively spliced to give a protein that acts as a vasodilator—two quite different gene products expressed from the same gene. Analysis of the human genome sequence has revealed that most human genes, 90% or more, are subject to alternative splicing.

**Figure 24.25**

The overall process of splicing. The pre-mRNA plus assorted snRNPs assemble and disassemble a spliceosome, which carries out the splicing reaction. The snRNPs are designated U1, U2, and so on. U1 is bound in step 1, which together with U2 binding (step 2) leads to a looped structure. Factors U4/6 and U5 then bind (step 3), and cleavage and transfer then occur (steps 4, 5). The spliceosome disassembles, releasing the ligated product (step 6) and the looped intron (step 7). This is degraded into small oligonucleotides (step 8).
All RNA is synthesized by the template-dependent copying of one DNA strand within a gene, catalyzed by RNA polymerase. RNA polymerases use 5’-ribonucleoside triphosphates as substrates, and they transcribe in a 5’ → 3’ direction. Bacteria synthesize all RNA classes with one polymerase, whereas eukaryotic cells have different polymerases—I, II, and III—for synthesis of ribosomal, messenger, and transfer RNA species, respectively.

Strand selection and duplex unwinding and rewinding are carried out by RNA polymerase in conjunction with other proteins. The enzyme binds at a promoter site, by formation of specific DNA–protein contacts, largely involving the enzyme’s σ subunit in bacteria and by a host of transcription factors and regulatory proteins in eukaryotes. Most initiations are abortive, but after a productive initiation involving different factors in bacteria and eukaryotes, elongation continues. In bacteria this is carried out by the core polymerase, αββ′ω. Transcription is highly progressive and is terminated by specific DNA sequences, sometimes in bacteria with the participation of ρ protein. Structural analyses have identified common features in bacterial and eukaryotic RNA polymerases, which have helped to reveal common mechanistic features.

Post-transcriptional RNA processing includes cutting pre-rRNA transcripts, which encode large and small rRNA components, and trimming tRNA precursors in both bacterial and eukaryotic cells with the aid of the ribozyme ribonuclease P, followed by modification of several nucleotides in each tRNA molecule and nontranscriptive addition of the 3’ terminal CCA sequence. Bacterial mRNAs undergo little if any post-transcriptional processing, whereas eukaryotic messages are extensively processed, with polyadenylation at the 3’ end, capping at the 5’ end with an inverted and modified guanine nucleotide residue, and splicing throughout the gene. Splicing is carried out by small nuclear ribonucleoprotein particles, guided by complementary base sequence interactions between splice sites and base sequences in the small nuclear RNA components. Alternative splicing is a process that expands the information content of a genome by directing quite different mRNA splicing patterns in different tissues and at different developmental stages.

1. Outline an experimental approach to determining the average chain growth rate for transcription in vivo. Chain growth rate is the number of nucleotides polymerized per minute per RNA chain.

2. Outline an experimental approach to determining the average RNA chain growth rate during transcription of a cloned gene in vitro.
3. Measurements of RNA chain growth rates are often led astray by the phenomenon of pausing, in which an RNA polymerase molecule stops transcription when it reaches certain sites, for intervals that may be as long as several seconds. How might pausing be detected?

4. Suppose you want to study the transcription in vitro of one particular gene in a DNA molecule that contains several genes and promoters. Without adding specific regulatory proteins, how might you stimulate transcription from the gene of interest relative to the transcription of the other genes on your DNA template? To make all of the complexes identical, you would like to arrest all transcriptional events at the same position on the DNA template before isolating the complex. How might you do this?

5. The tac promoter, an artificial promoter made from portions of the trp and lacUV5 promoters, has been introduced into a plasmid. It is a hybrid of the lac and trp (tryptophan) promoters, containing the −35 region of one and the −10 region of the other. This promoter directs transcription initiation more efficiently than either the trp or lac promoters. Why?

6. Explain the basis for the following statement: Transcription of two genes on a plasmid can occur without the concomitant action of a topoisomerase, but only if those two genes are oriented in opposite directions.

7. Some years ago, it was suggested that the function of the poly(A) tail on a eukaryotic mRNA message may be to “ticket” the message. That is, each time the message is used, one or more residues is removed, and the message is degraded after the tail is shortened below a critical length. Suggest an experiment to test this hypothesis.

8. For the original detection of DNA–RNA hybrid molecules, as described on page 733, the DNA–RNA hybrid was detected in a CsCl equilibrium gradient. Why are RNA and DNA–RNA hybrids denser than double-stranded DNA?

9. Shown below is an R loop prepared for electron microscopy by annealing a purified eukaryotic messenger RNA with DNA from a genomic clone containing the full-length gene corresponding to the mRNA.

(a) How many exons does the gene contain? How many introns?
(b) Where in this structure would you expect to find a 5′,5′′-internucleotide bond? Where would you expect to find a polyadenylic acid sequence?

10. Introns in protein-coding genes of some eukaryotes are rarely shorter than 65 nucleotides long. What might be a rationale for this limitation?

11. Heparin is a polyanionic polysaccharide that blocks initiation by RNA polymerase by virtue of its binding to double-stranded DNA. But heparin inhibits only when added before the onset of transcription, and not if added after transcription begins. Explain this difference.

12. Estimate the time needed for E. coli RNA polymerase at 37 °C to transcribe the entire gene for a 50-kilodalton protein. What assumption or assumptions must be made for this estimate to be accurate?

13. Describe how RNA polymerase backtracking could function to increase the fidelity of transcription.

14. Is RNA polymerase saturated with substrates in vivo? Describe experiments that might indicate whether RNA polymerase is operating at Vmax with respect to its nucleotide substrates.

15. As discussed in the text, promoters were originally identified as consensus sequences upstream from transcriptional start sites. What additional evidence might support the assignment of these sequences as parts of promoters?

16. In this chapter and elsewhere, we have described two types of ultracentrifugation experiments—sucrose-gradient centrifugation and equilibrium-density-gradient centrifugation. Briefly discuss these procedures with respect to the physical bases on which molecular species are separated in each method, the kinds of isotopic compounds used in the analysis, and the biological processes that have been or can be analyzed with each.

17. About 98% of the E. coli genome codes for proteins, yet mRNA, the template for protein synthesis, comprises only about 2% of the total RNA in the cell. Explain this apparent discrepancy.

REFERENCES


The finding in the 1960s that single-stranded DNA binds irreversibly to membrane filters, and the development of recombinant DNA technology in the 1970s, led to a number of techniques for analyzing gene expression (i.e., measuring the levels of transcripts of particular genes in living cells). RNA could be radiolabeled in vivo and hybridized to gene-specific DNA—a cloned gene or a restriction fragment—and the bound radioactivity analyzed by autoradiography or in a liquid scintillation counter. Several techniques, such as Northern analysis, were based on these developments. Northern analysis is comparable to Southern analysis, except that in Northern analysis RNA is hybridized to DNA restriction fragments separated by gel electrophoresis and then immobilized on a filter. However, such approaches allow the analysis of just one or a few genes in each experiment. With the availability of complete genome sequences, it became desirable to analyze levels of transcripts from many genes in a single experiment, that is, patterns of gene expression, which could be compared under different physiological conditions. Microarray technology makes this kind of analysis possible.

In a microarray experiment, minute amounts of gene-specific DNAs—usually several thousand—are immobilized on a substrate, such as glass or a membrane filter. The gene-specific DNAs are either cloned cDNAs or synthetic oligonucleotides. Using robotic technology, the investigator "prints" single-stranded DNAs onto the substrate, which may be a microscope slide, suitably coated to bind the applied DNAs. The DNAs are printed as a large array, which allows the investigator to identify each gene from its position on the array. The DNAs are fixed irreversibly on the substrate, so that the "DNA chip" can be used repeatedly by stripping the annealed RNA targets off the chip after each experiment.

Typically, a microarray experiment involves the comparison of gene-expression profiles under different conditions—comparing a tumor with the tissue of origin, for example, or comparing a hormone-stimulated tissue with unstimulated tissue (see Figure 24A.1). The investigator wishes to learn which genes are activated under the conditions being analyzed and which are repressed. Total mRNA is...
isolated from each tissue or cell culture and converted to a population of cDNAs using reverse transcriptase. During the enzymatic synthesis of the cDNAs, one of the deoxyribonucleoside triphosphates is tagged with a fluorescent dye. Typically, the reference sample is labeled with a red fluorophore, and the test sample is labeled with a green fluorophore. After cDNA synthesis is complete, the two samples are mixed and subjected to annealing conditions in the presence of the microarray. Unhybridized cDNAs are washed off, and the array is then scanned. Scanning at wavelengths corresponding to emission maxima of the fluorophores reveals which transcripts are more abundant in the test than in the reference (more green fluorescence) and which are less abundant (more red fluorescence). Analysis of the image reveals which genes were stimulated and which repressed under the conditions being tested. An example of a microarray analysis was shown in Figure 1.11.

Microarray technology has numerous applications in addition to measuring patterns of gene expression. For example, by using arrayed oligonucleotides representing different mutant forms of a gene of interest, we can carry out DNA–DNA hybridization on the gene chip and identify mutations or single-nucleotide polymorphisms in biological samples.

References


So far the techniques we have discussed for identifying and characterizing binding sites for proteins such as RNA polymerase and repressors have involved single binding sites, whereas regulatory proteins such as nuclear hormone receptors act at multiple binding sites. Moreover, techniques such as footprinting are carried out in vitro, although our main interest is characterizing attachment sites for DNA-binding proteins in intact cells. Chromatin immunoprecipitation (ChIP) makes it possible to identify in vivo binding sites on a genome-wide basis.

The principle of ChIP is that any DNA-binding protein can be covalently attached to its DNA binding site(s) in vivo by use of a crosslinking reagent that can penetrate cell membranes and react covalently with both protein and DNA in a reversible manner. Formaldehyde is most commonly used, as shown in Figure 24B.1. After whole cells are treated with formaldehyde, chromatin is isolated and subjected to sonic oscillation under conditions that reduce the length of each DNA molecule to fragments several hundred base pairs long. The mixture is treated with antibody to the protein of interest, and the immunoprecipitated DNA–protein complexes are collected. At this point the crosslinks are broken, and the DNA that was precipitated along with the protein is subjected to sequence analysis. Originally this was done most often by PCR amplification of the DNA followed by conventional sequence analysis. This approach, however, makes it possible only to analyze known and suspected DNA sequences. An alternate approach involves cloning all DNA fragments in the mixture and then using PCR primers corresponding to flanking sequences on the vector, followed by sequence analysis of each clone. With the advent of microarray technology (Tools of Biochemistry 24A) it became possible to screen the DNA fragments against a DNA microarray containing hundreds or thousands of DNA sequences. This technique is called ChIP-chip, because the immunoprecipitated DNA fragments are identified on a gene chip. A still more recent innovation, called ChIP-seq, sequences all of the immunoprecipitated DNA in parallel using a next-generation sequencing technology that permits simultaneous sequence analysis of hundreds or thousands of DNA molecules.

References

Chromatin immunoprecipitation. A transcription factor (TF) is crosslinked to the DNA sites to which it binds in chromatin. After fragmentation and immunoprecipitation using antibody to TF, the TF-bound DNA fragments are identified either by microarray analysis, in which the DNA is bound to a red fluorophore and subjected to hybridization analysis with an array of genomic fragments (ChIP-chip), or subjected to massive parallel sequence analysis (ChIP-seq). Data from B. J. Venters and B. F. Pugh (2009) How eukaryotic genes are transcribed, Critical Reviews in Biochemistry and Molecular Biology 44:117–141.