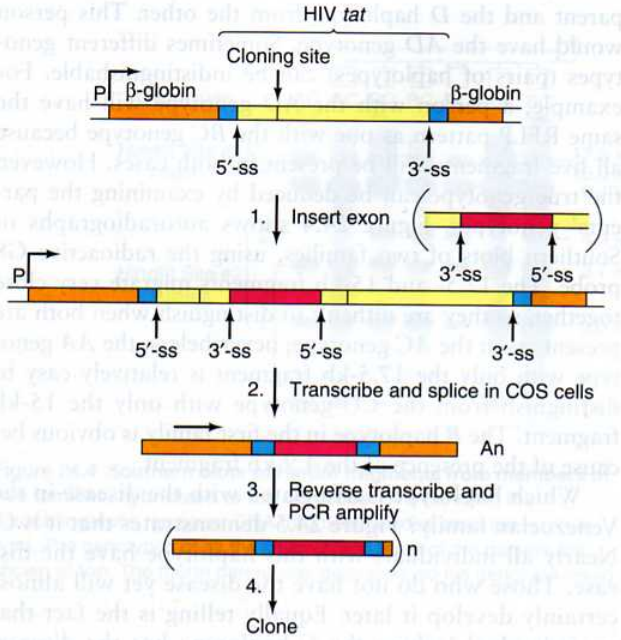


igure 24.2 Exon trapping. Begin with a cloning vector, such as SPL1, shown here in slightly simplified form. This vector has an SV40 romoter (P), which drives expression of a hybrid gene containing the abbit β-globin gene (orange), interrupted by part of the HIV *tat* gene, /hich includes two exon fragments (blue) surrounding an intron ellow). The exon–intron borders contain 5′- and 3′-splice sites (ss). he *tat* intron contains a cloning site, into which random DNA agments can be inserted. In step 1, an exon (red) has been inserted, anked by parts of its own introns, and its own 5′- and 3′-splice sites. step 2, insert this construct into COS cells, where it can be anscribed and then the transcript can be spliced. Note that the oreign exon (red) has been retained in the spliced transcript, because had its own splice sites. Finally (steps 3 and 4), subject the anscripts to reverse transcription and PCR amplification, with rimers indicated by the arrows. This gives many copies of a DNA agment containing the foreign exon, which can now be cloned and xamined. Note that a non-exon will not have splice sites and will herefore be spliced out of the transcript along with the intron. It will ot survive to be amplified in step 3, so one does not waste time udying it.

eneticists can scan large regions of DNA for "islands" of tes that could be cut with *Hpa*II in a "sea" of other DNA equences that could not be cut. Such a site is called a **CpG land,** or an **HTF island** because it yields *Hpa*II tiny agments.

SUMMARY Positional cloning begins with mapping studies (Chapter 1) to pin down the location of the gene of interest to a reasonably small region of DNA. Mapping depends on a set of landmarks to which the position of a gene can be related. Sometimes such landmarks are genes, but more often they are RFLPs—sites at which the lengths of restriction fragments generated by a given restriction enzyme vary from one individual to another. Several methods are available for identifying the genes in a large region of unsequenced DNA. One of these is the exon trap, which uses a special vector to help clone exons only. Another is to use methylation-sensitive restriction enzymes to search for CpG islands—DNA regions containing unmethylated CpG sequences.

Identifying the Gene Mutated in a Human Disease

Let us conclude this section with a classic example of positional cloning: pinpointing the gene for Huntington disease.

Huntington disease (HD) is a progressive nerve disorder. It begins almost imperceptibly with small tics and clumsiness. Over a period of years, these symptoms intensify and are accompanied by emotional disturbances. Nancy Wexler, an HD researcher, describes the advanced disease as follows: "The entire body is encompassed by adventitious movements. The trunk is writhing and the face is twisting. The full-fledged Huntington patient is very dramatic to look at." Finally, after 10–20 years, the patient dies.

Huntington disease is controlled by a single dominant gene. Therefore, a child of an HD patient has a 50:50 chance of being affected. People who have the disease could avoid passing it on by not having children, except that the first symptoms usually do not appear until after the childbearing years.

Because they did not know the nature of the product of the HD gene (*HD*), geneticists could not look for the gene directly. The next best approach was to look for a gene or other marker that is tightly linked to *HD*. Michael Conneally and his colleagues spent more than a decade trying to find such a linked gene, but with no success.

In their attempt to find a genetic marker linked to *HD*, Wexler, Conneally, and James Gusella turned next to RFLPs. They were fortunate to have a very large family to study. Living around Lake Maracaibo in Venezuela is a family whose members have suffered from HD since the early nineteenth century. The first member of the family to be so afflicted was a woman whose father, presumably a European, carried the defective gene. So the pedigree of this family can be traced through seven generations, and the number of individuals is unusually large: It is not uncommon for a family to have 15–18 children.

Gusella and colleagues knew they might have to test hundreds of probes to detect a RFLP linked to *HD*, but they were amazingly lucky. Among the first dozen probes they tried, they found one (called G8) that detected a RFLP

that is very tightly linked to *HD* in the Venezuelan family. Figure 24.3 shows the locations of *Hin*dIII sites in the stretch of DNA that hybridizes to the probe. We can see seven sites in all, but only five of these are found in all family members. The other two, marked with asterisks and numbered 1 and 2, may or may not be present. These latter two sites are therefore polymorphic, or variable.

Let us see how the presence or absence of these two restriction sites gives rise to a RFLP. If site 1 is absent, a single fragment 17.5 kb long will be produced. However, if site 1 is present, the 17.5-kb fragment will be cut into two pieces having lengths of 15 kb and 2.5 kb, respectively. Only the 15-kb band will show up on the autoradiograph because the 2.5-kb fragment lies outside the region that hybridizes to the G8 probe. If site 2 is absent, a 4.9-kb fragment will be produced. On the other hand, if site 2 is present, the 4.9-kb fragment will be subdivided into a 3.7-kb fragment and a 1.2-kb fragment.

There are four possible **haplotypes** (clusters of alleles on a single chromosome) with respect to these two polymorphic *Hin*dIII sites, and they have been labeled *A–D:*

Haplotype	Site 1	Site 2	Fragments Observed
A	Absent	Present	17.5; 3.7; 1.2
B	Absent	Absent	17.5; 4.9
C	Present	Present	15.0; 3.7; 1.2
D	Present	Absent	15.0; 4.9

The term *haplotype* is a contraction of *haploid genotype,* which emphasizes that each member of the family will inherit two haplotypes, one from each parent. For example, an individual might inherit the *A* haplotype from one parent and the *D* haplotype from the other. This person would have the *AD* genotype. Sometimes different genotypes (pairs of haplotypes) can be indistinguishable. For example, a person with the *AD* genotype will have the same RFLP pattern as one with the *BC* genotype because all five fragments will be present in both cases. However, the true genotype can be deduced by examining the parents' genotypes. Figure 24.4 shows autoradiographs of Southern blots of two families, using the radioactive G8 probe. The 17.5- and 15-kb fragments migrate very close together, so they are difficult to distinguish when both are present, as in the *AC* genotype; nevertheless, the *AA* genotype with only the 17.5-kb fragment is relatively easy to distinguish from the *CC* genotype with only the 15-kb fragment. The *B* haplotype in the first family is obvious because of the presence of the 4.9-kb fragment.

Which haplotype is associated with the disease in the Venezuelan family? Figure 24.5 demonstrates that it is *C.* Nearly all individuals with this haplotype have the disease. Those who do not have the disease yet will almost certainly develop it later. Equally telling is the fact that no individual lacking the *C* haplotype has the disease. Thus, this is a very accurate way of predicting whether a member of this family is carrying the Huntington disease gene. A similar study of an American family showed that, in this family, the *A* haplotype was linked with the disease. Therefore, each family varies in the haplotype associated with the disease, but within a family, the linkage between the RFLP site and *HD* is so close that recombination between these sites is very rare. Thus we see that a RFLP can be used as a genetic marker for mapping, just as if it were a gene.

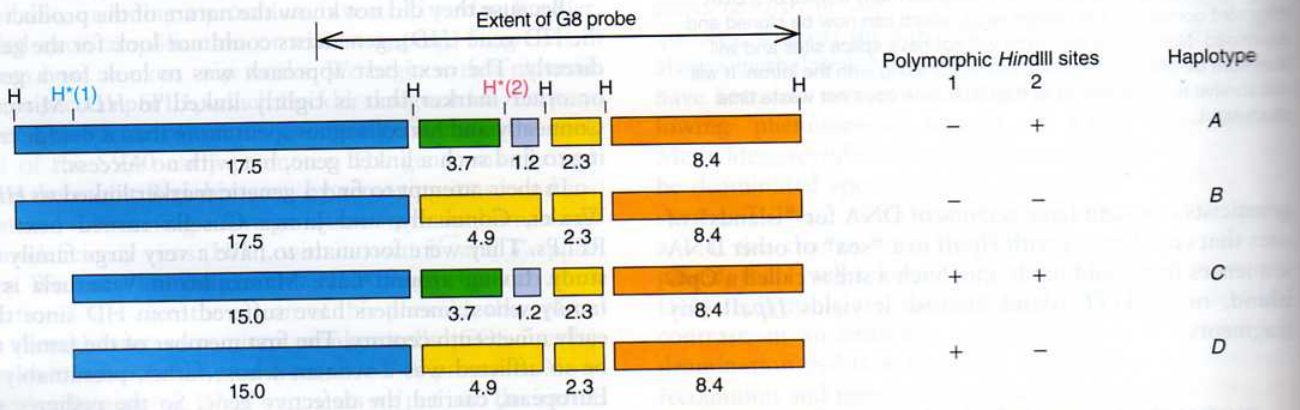


Figure 24.3 The RFLP associated with the Huntington disease gene. The *Hin*dIII sites in the region that hybridizes to the G8 probe are shown. The families studied show polymorphisms in two of these sites, marked with an asterisk and numbered 1 (blue) and 2 (red). Presence of site 1 results in a 15-kb fragment plus a 2.5-kb fragment that is not detected because it lies outside the region that hybridizes to the G8 probe. Absence of this site results in a 17.5-kb fragment. Presence of site 2 results in two fragments of 3.7 and 1.2 kb. Absence of this site results in a 4.9-kb fragment. Four haplotypes (*A–D*) result from the four combinations of presence or absence of these two sites. These are listed at right, beside a list of polymorphic *Hin*dIII sites and a diagram of the *Hin*dIII restriction fragments detected by the G8 probe for each haplotype. For example, haplotype *A* lacks site 1 but has site 2. As a result, *Hin*dIII fragments of 17.5, 3.7, and 1.2 are produced. The 2.3- and 8.4-kb fragments are also detected by the probe, but we ignore them because they are common to all four haplotypes.

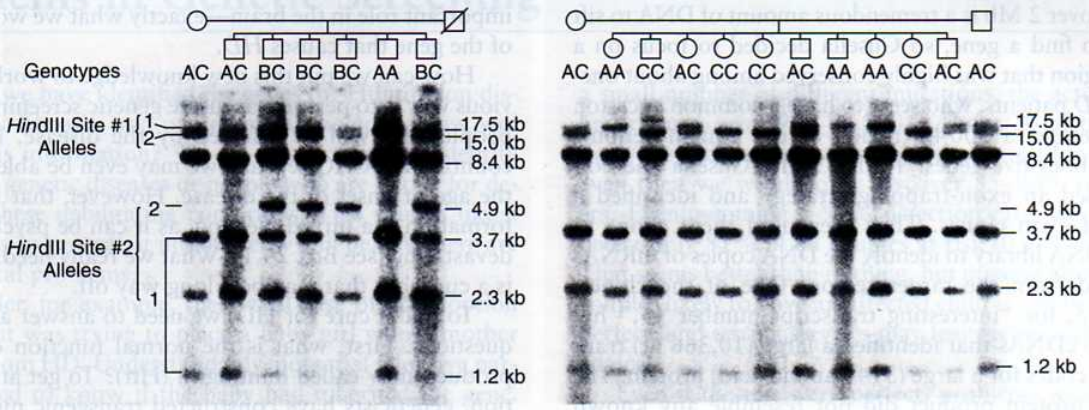


Figure 24.4 Southern blots of *Hin*dIII fragments from members of two families, hybridized to the G8 probe. The bands in the autoradiographs represent DNA fragments whose sizes are listed at right. The genotypes of all the children and three of the parents are shown at top. The fourth parent was deceased, so his genotype could not be determined. (*Source:* Gusella, J.F., N.S. Wexler, P.M. Conneally, S.L. Naylor, M.A. Anderson, R.E. Tauzi, et al., A polymorphic DNA marker genetically linked to Huntington's disease. *Nature* 306:236. Copyright © 1983 Macmillan Magazines Limited.)

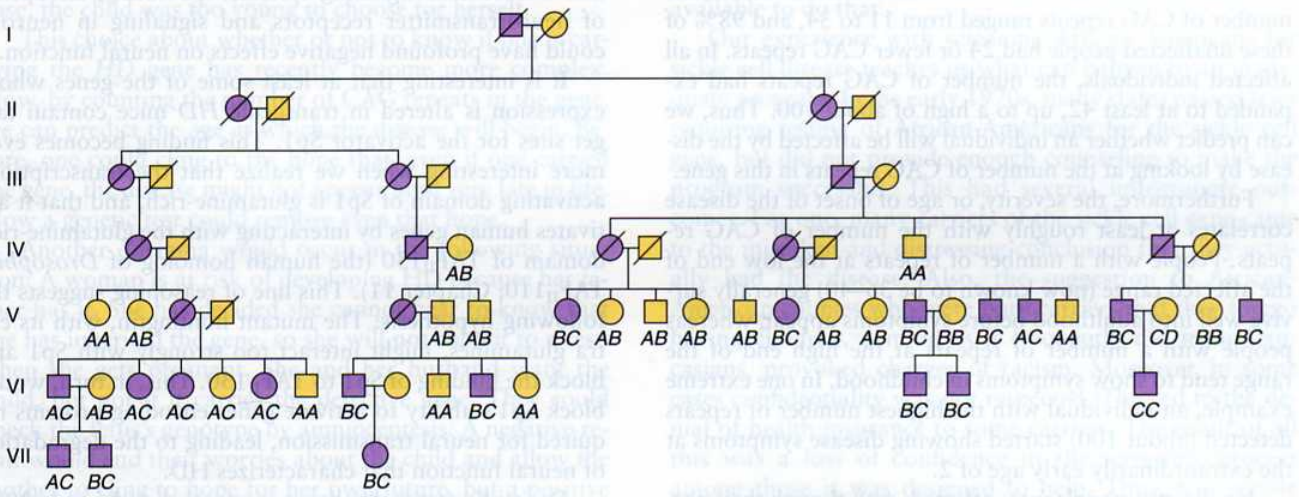


Figure 24.5 Pedigree of the large Venezuelan family with Huntington disease. Family members with confirmed disease are represented by purple symbols. Notice that most of the individuals with the *C* haplotype already have the disease, and that no sufferers of the disease lack the *C* haplotype. Thus, the *C* haplotype is strongly associated with the disease, and the corresponding RFLP is tightly linked to the Huntington disease gene.

Finding linkage between *HD* and the DNA region that hybridizes to the G8 probe also allowed Gusella and colleagues to locate *HD* to chromosome 4. They did this by making mouse–human hybrid cell lines, each containing only a few human chromosomes. They then prepared DNA from each of these lines and hybridized it to the radioactive G8 probe. Only the cell lines having chromosome 4 hybridized; the presence or absence of all other chromosomes did not matter. Therefore, human chromosome 4 carries *HD*.

At this point, the *HD* mapping team's luck ran out. One long detour arose from a mapping study that indicated the gene lay far out at the end of chromosome 4. This made the search much more difficult because the tip of the chromosome is a genetic wasteland, full of repetitive sequences, and apparently devoid of genes. Finally, after wandering for years in what he called a genetic "junkyard," Gusella and his group turned their attention to a more promising region. Some mapping work suggested that *HD* resided, not at the tip of the chromosome, but in a 2.2-Mb region several

information from genomic and proteomic research can be used.

Finally, we will introduce **bioinformatics,** the discipline concerned with managing and using the vast stores of data that come from genomic, proteomic, and other massive biological studies.

24.1 Positional Cloning: An Introduction to Genomics

Before we examine the techniques of genomic research, let us consider one of the important uses of genomic information: **positional cloning,** which is one method for the discovery of the genes involved in genetic traits. In humans, this frequently involves the identification of genes that govern genetic diseases. We will begin by considering an example of positional cloning that was done before the genomic era: finding the gene whose malfunction causes Huntington disease in humans. We will see that much of the effort went into narrowing down the region in which to look for the faulty gene. One reason for all this effort was to avoid having to sequence a huge chunk of DNA. Nowadays, that is not a problem because the sequencing has already been done. Nevertheless, this example serves as a good introduction to genomics for several reasons: It illustrates the principle of positional cloning, which is still a major use of genomic information; it shows how difficult positional cloning was in the absence of genomic information; and it is a heroic story that still deserves to be told.

Classical Tools of Positional Cloning

Geneticists seeking the genes responsible for human genetic disorders frequently face a problem: They do not know the identity of the defective protein, so they are looking for a gene without knowing its function. Thus, they have to identify the gene by finding its position on the human genetic map, and this process therefore has come to be called positional cloning.

The strategy of positional cloning begins with the study of a family or families afflicted with the disorder, with the goal of finding one or more markers that are tightly linked to the "disease gene," that is, the gene which, when mutated, causes the disease. Frequently, these markers are not genes, but stretches of DNA whose pattern of cleavage by restriction enzymes or other physical attributes vary from one individual to another.

Because the position of the marker is known, the disease gene can be pinned down to a relatively small region of the genome. However, that "relatively small" region usually contains about a million base pairs, so the job is not over. The next step is to search through the million or so base pairs to find a gene that is the likely culprit. Several tools have traditionally been used in the search, and we will describe two here. These are: (1) finding exons with exon traps; and (2) locating the CpG islands that tend to be associated with genes. We will see how these tools have been used as we discuss our example in the next section of this chapter.

Restriction Fragment Length Polymorphisms In the late twentieth century, we knew the locations of relatively few human genes, so the likelihood of finding one of these close to a new gene we were trying to map was small. Another approach, which does not depend on finding linkage with a known gene, is to establish linkage with an "anonymous" stretch of DNA that may not even contain any genes. We can recognize such a piece of DNA by its pattern of cleavage by restriction enzymes.

Because each person differs genetically from every other, the sequences of their DNAs will differ a little bit, as will the pattern of cutting by restriction enzymes. Consider the restriction enzyme *Hin*dIII, which recognizes the sequence AAGCTT. One individual may have three such sites separated by 4 and 2 kb, respectively, in a given region of a chromosome (Figure 24.1). Another individual may lack the middle site but have the other two, which are 6 kb apart. This means that if we cut the first person's DNA with *Hin*dIII, we will produce two fragments, 2 kb and 4 kb long, respectively. The second person's DNA will yield a 6-kb fragment instead. In other words, we are dealing with a **restriction fragment length polymorphism (RFLP).** Polymorphism means that a genetic locus has different forms, or alleles (Chapter 1), so this clumsy term simply means that cutting the DNA from any two individuals with a restriction enzyme may yield fragments of different lengths. The abbreviated term, RFLP, is usually pronounced "rifflip."

How do we go about looking for a RFLP? Clearly, we cannot analyze the whole human genome at once. It contains hundreds of thousands of cleavage sites for a typical restriction enzyme, so each time we cut the whole genome with such an enzyme, we release hundreds of thousands of fragments. No one would relish sorting through that morass for subtle differences between individuals.

Fortunately, there is an easier way. With a Southern blot (Chapter 5) one can highlight small portions of the total genome with various probes, so any differences are easy to see. However, there is a catch. Because each labeled probe hybridizes only to a small fraction of the total human DNA, the chances are very poor that any given one will reveal a RFLP linked to the gene of interest. We may have to screen thousands of probes before we find the right one. As laborious as it is, this procedure at least provides a starting point, and it has been a key to finding the genes responsible for several genetic diseases.

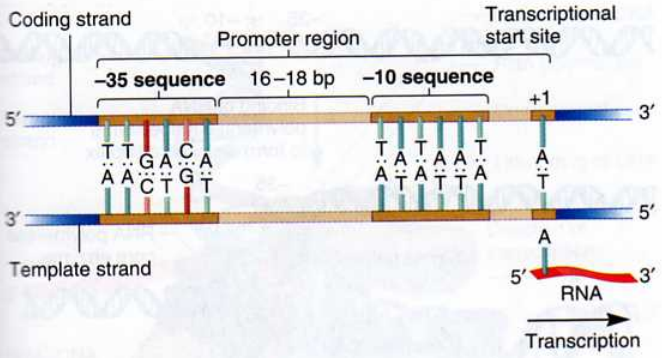


FIGURE 12.3 The conventional numbering system of promoters. The first nucleotide that acts as a template for transcription is designated +1. The numbering of nucleotides to the left of this spot is in a negative direction, while the numbering to the right is in a positive direction. For example, the nucleotide that is immediately to the left of the +1 nucleotide is numbered −1, and the nucleotide to the right of the +1 nucleotide is numbered +2. There is no zero nucleotide in this numbering system. In many bacterial promoters, sequence elements at the −35 and −10 regions play a key role in promoting transcription.

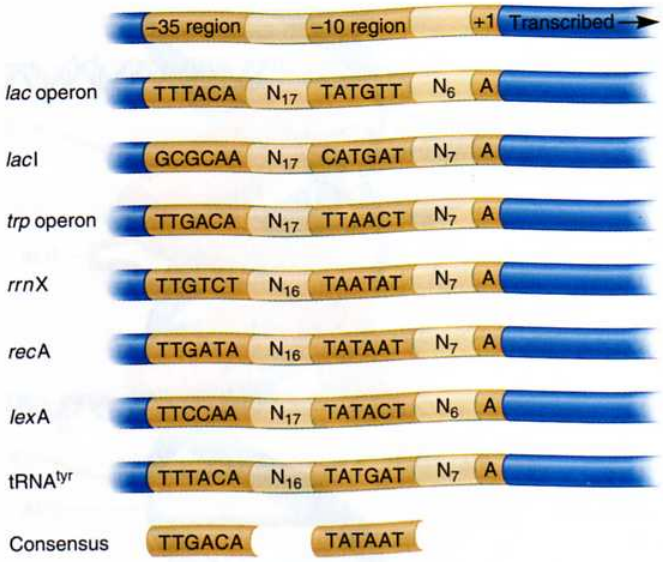


FIGURE 12.4 Examples of −35 and −10 sequences within a variety of bacterial promoters. This figure shows the −35 and −10 sequences for seven different bacterial and bacteriophage promoters. The consensus sequence is shown at the *bottom*. The spacer regions contain the designated number of nucleotides between the −35 and −10 region or between the −10 region and the transcriptional start site. For example, N_{17} means that there are 17 nucleotides between the end of the −35 region and the beginning of the −10 region.

Although the promoter may encompass a region that is several dozen nucleotides in length, there are short **sequence elements** that are particularly critical for promoter recognition. By comparing the sequence of DNA bases within many promoters, researchers have learned that certain sequences of bases are necessary to create a functional promoter. In many promoters found in *E. coli* and similar species, two sequence elements are important. These are located at approximately the −35 and −10 sites in the promoter region (fig. 12.3). The sequence at the −35 region is 5′–TTGACA–3′, and the one at the −10 region is 5′–TATAAT–3′. The TATAAT sequence is sometimes called the **Pribnow box** after David Pribnow, who initially discovered it in 1975.

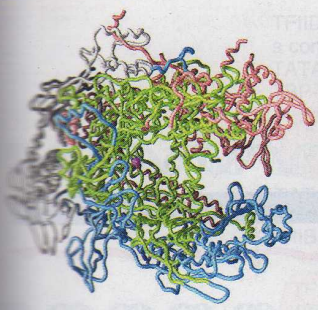
The sequences at the −35 and −10 sites can vary among different genes. For example, figure 12.4 illustrates the sequences found in several different *E. coli* promoters. The most commonly occurring bases within a sequence element form the **consensus sequence.** It is usually the sequence that is most efficiently recognized. For many bacterial genes, there is a good correlation between the rate of RNA transcription and the degree to which the −35 and −10 regions agree with their consensus sequences.

Bacterial Transcription Is Initiated When RNA Polymerase Holoenzyme Binds at a Promoter Sequence

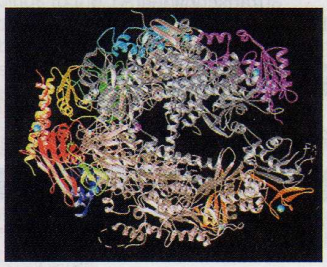
Thus far, we have considered the DNA sequences that constitute a functional promoter. Let's now turn our attention to a consideration of the proteins that recognize those sequences and carry out the transcription process. The enzyme that catalyzes the synthesis of RNA is **RNA polymerase.** In *E. coli,* the **core enzyme** is composed of four subunits, $\alpha_2\beta\beta'$. The association of a fifth subunit, **sigma factor,** with the core enzyme is referred to as RNA polymerase **holoenzyme.** The different subunits within the holoenzyme play distinct functional roles. The two α subunits are important in the proper assembly of the holoenzyme and in the process of binding to DNA. The β and β' subunits are also needed for binding to the DNA and are critical in the catalytic synthesis of RNA. The holoenzyme is required to initiate transcription; the primary role of sigma factor is to recognize the promoter. Proteins, such as sigma factor, that influence the function of RNA polymerase are known as **transcription factors.** The core enzyme is necessary for RNA synthesis.

After RNA polymerase holoenzyme is assembled into its five subunits, it binds loosely to the DNA. It then scans along the DNA much as a train rolls down the tracks. When it encounters a promoter region, sigma factor recognizes the promoter elements at both the −35 and −10 positions. A region within the sigma factor protein that contains a **helix-turn-helix structure** is involved in a tighter binding to the DNA. Alpha helices within the protein can fit into the major groove of the DNA double helix and form hydrogen bonds with the bases. This phenomenon is shown in figure 12.5. Hydrogen bonding occurs between nucleotides in the −35 and −10 regions of the promoter and amino acid side chains in the helix-turn-helix structure of sigma factor.

As shown in figure 12.6, the process of transcription is initiated when sigma factor within the holoenzyme has bound to the promoter region to form the **closed complex.** For transcription to begin, the double-stranded DNA must then be unwound into an open complex. This unwinding first occurs at the TATAAT box in the −10 region, which contains mostly AT base

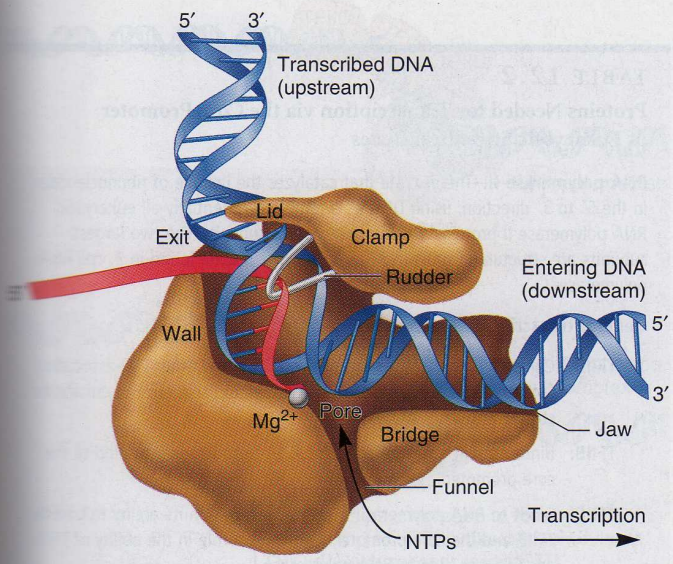


Structure of a bacterial RNA polymerase



Structure of a eukaryotic RNA polymerase II

(a)



(b) Schematic structure of RNA polymerase

FIGURE 12.10 Structure and molecular function of RNA polymerase. (**a**) A comparison of the crystal structures of a bacterial RNA polymerase (*left*) to a eukaryotic RNA polymerase II (*right*). The bacterial enzyme is from *Thermus aquaticus*. The eukaryotic enzyme is from *Saccharomyces cerevisiae*. (**b**) A molecular mechanism for transcription based on the crystal structure. In this diagram, the direction of transcription is from left to right. The double-stranded DNA enters the polymerase along a bridge surface that is between the jaw and clamp. At a region termed the wall, the DNA-RNA hybrid is forced to make a right-angle turn, which enables nucleotides to bind to the template strand. Mg^{2+} is located at the catalytic site. Nucleoside triphosphates (NTPs) enter the catalytic site via a funnel and pore region and bind to the template DNA. At the catalytic site, the nucleotides are covalently attached to the 3′ end of the RNA. As RNA polymerase slides down the template, a small region of the protein termed the rudder separates the RNA-DNA hybrid. The single-stranded RNA then exits under a small lid.

would bind to the template DNA and then be covalently attached to the 3′ end. As RNA polymerase slides down the template, a rudder, which is about 9 bp away from the 3′ end of the RNA, forces the RNA-DNA hybrid apart. The single-stranded RNA then exits under a small lid.

Eukaryotic Structural Genes Have a Core Promoter and Regulatory Elements

In eukaryotes, the promoter sequence is more variable and often more complex than that found in bacteria. For structural genes, at least three features are found in most promoters: a **transcriptional start site,** a **TATA box,** and **regulatory elements.** Figure 12.11 shows a common pattern of sequences found within the promoters of eukaryotic structural genes. The **core promoter** is relatively short. It consists of a TATAAAA sequence called the TATA box and a transcriptional start site. The TATA box, which is usually about 25 bp upstream from a transcriptional start site, is important in determining the precise starting point for transcription. If it is missing from the core promoter, the transcription starting point becomes undefined, and transcription may start at a variety of different locations. The core promoter, by itself, produces a low level of transcription. This is termed **basal transcription.**

Regulatory elements affect the ability of RNA polymerase to recognize the core promoter and begin the process of transcription. There are two categories of regulatory elements. Activating sequences, known as **enhancers,** are needed to stimulate transcription. In the absence of enhancer sequences, most eukaryotic genes have very low levels of basal transcription. Under certain conditions, it is necessary to prevent transcription of a given gene. This occurs via repressing sequences, known as **silencers,** which inhibit transcription. In chapter 12, we will briefly consider the locations of regulatory elements relative to the core promoter. The functions of eukaryotic regulatory elements, and the **regulatory transcription factors** that bind to them, are examined in greater detail in chapter 15. As seen in figure 12.11, a common location for regulatory elements is the −50 to −100 region. However, the locations of regulatory elements are quite variable among different eukaryotic genes. Such elements can be far away from the core promoter yet exert strong effects on the ability of RNA polymerase to initiate transcription.

With regard to terminology, there are two ways to view the relative locations of factors that control the expression of genes. DNA sequences such as the TATA box, enhancers, and silencers exert their effects only over a nearby gene. They are called ***cis*-acting elements.** The term *cis* comes from chemistry nomenclature meaning "next to." By comparison, the regulatory proteins that bind to such elements are called ***trans*-acting factors,** the term *trans* meaning "across from." The proteins that control the expression of a particular gene are also encoded by genes (i.e., regulatory genes), but such regulatory genes may be far away from the genes they control. Nevertheless, when a regulatory gene encoding a *trans*-acting factor is expressed, the protein that is made can diffuse throughout the cell and bind to its appropriate *cis*-acting element. Let's now turn our attention to the function of proteins that bind to *cis*-acting elements.

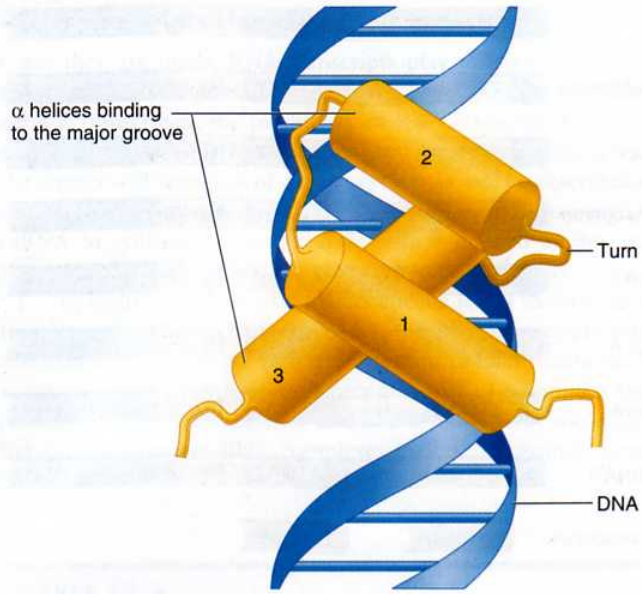


FIGURE 12.5 The binding of a transcription factor protein to the DNA double helix. In this example, the protein contains a helix-turn-helix motif. Two α helices of the protein (labeled *2* and *3*) can fit within the major groove of the DNA. Amino acids within the α helices form hydrogen bonds with the bases in the DNA.

pairs, as shown earlier in figure 12.4. As you may recall from chapter 9, AT base pairs form only two hydrogen bonds, whereas GC pairs form three. Therefore, it is easier to separate DNA in an AT-rich region, since fewer hydrogen bonds must be broken. A short strand of RNA is made within the open complex and then sigma factor is released from the core enzyme. The release of sigma factor marks the transition to the elongation phase of transcription. The core enzyme may now slide down the DNA to synthesize a strand of RNA.

The RNA Transcript Is Synthesized During the Elongation Stage

After the initiation stage of transcription is completed, the RNA transcript is made in the elongation stage of transcription. During the synthesis of the RNA transcript, RNA polymerase moves along the DNA, causing it to unwind (fig. 12.7). The DNA strand that is used as a template for RNA synthesis is called the **template** or **noncoding strand.** The opposite DNA strand is called the **coding strand;** it has the same sequence as the RNA transcript except that T in the DNA corresponds to U in the RNA. Within a given gene, only the template strand is used for RNA synthesis while the coding strand is never used. As it moves down the DNA, the open complex formed by the action of RNA polymerase is approximately 17 bp long. On average, the rate of RNA synthesis is about 43 nucleotides per second! Behind the open complex, the DNA rewinds back into a double helix.

As described in figure 12.7, the chemistry of transcription by RNA polymerase is similar to the synthesis of DNA via DNA

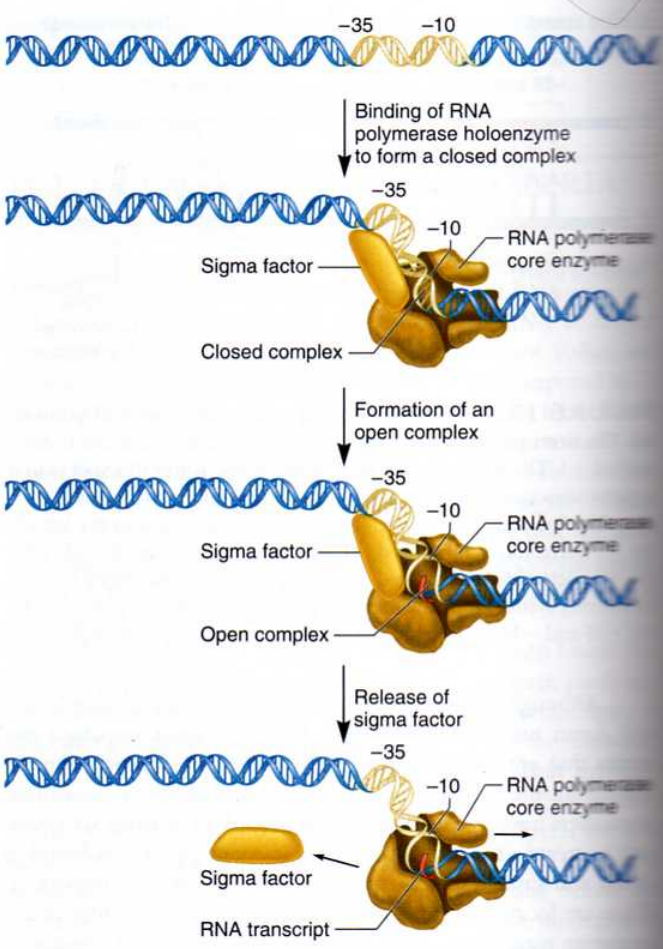
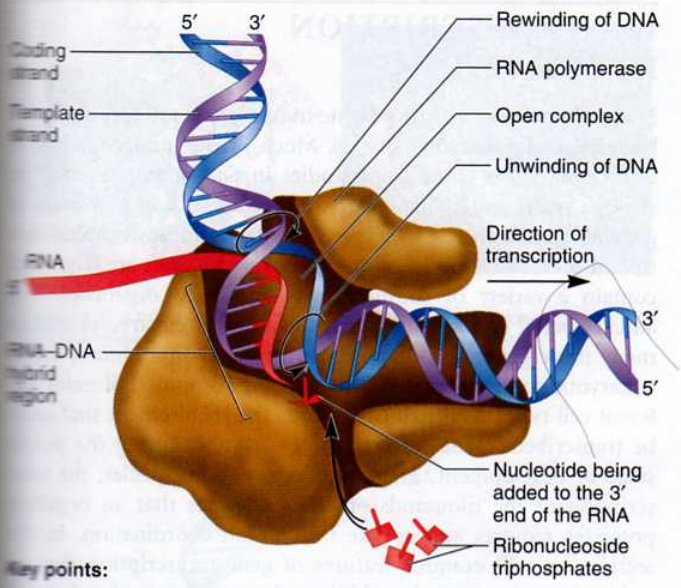


FIGURE 12.6 The initiation stage of transcription in bacteria. The sigma factor subunit of the RNA polymerase holoenzyme recognizes the −35 and −10 regions of the promoter. The DNA unwinds in the −10 region to form an open complex. Sigma factor then dissociates from the holoenzyme, and the RNA polymerase core enzyme can proceed down the DNA to transcribe RNA.

polymerase, which was discussed in chapter 11. RNA polymerase always connects nucleotides in the 5′ to 3′ direction. During this process, RNA polymerase catalyzes the formation of a bond between the 5′ phosphate group on one nucleotide and the 3′ –OH group on the previous nucleotide. The complementarity rule is similar to the AT/GC rule, except that uracil substitutes for thymine in the RNA. In other words, RNA synthesis obeys an $A_{RNA}T_{DNA}/U_{RNA}A_{DNA}/G_{RNA}C_{DNA}/C_{RNA}G_{DNA}$ rule.

Transcription Is Terminated by Either an RNA-Binding Protein or an Intrinsic Terminator

The end of RNA synthesis is referred to as termination. Prior to termination, the hydrogen bonding between the DNA and RNA within the open complex is of central importance in preventing dissociation of the RNA polymerase from the template strand.



Key points:

- RNA polymerase slides along the DNA, creating an open complex as it moves.
- The DNA strand known as the template strand is used to make a complementary copy of RNA as an RNA–DNA hybrid.
- The RNA is synthesized in a 5′ to 3′ direction using ribonucleoside triphosphates as precursors. Pyrophosphate is released (not shown).
- The complementarity rule is the same as the AT/GC rule except that U is substituted for T in the RNA.

FIGURE 12.7 Synthesis of the RNA transcript.

Termination occurs when this short RNA-DNA hybrid region is forced to separate, thereby releasing the newly made RNA transcript as well as RNA polymerase. In *E. coli*, two different mechanisms for termination have been identified. For certain genes, a protein known as **ρ (rho)** is responsible for terminating transcription, a mechanism called **ρ-dependent termination.** For other genes, termination does not require ρ. This is referred to as **ρ-independent termination.** Both mechanisms will be described here.

In ρ-dependent termination, a sequence near the 3′ end of the newly made RNA called the *rut* site (for rho utilization site) acts as a recognition site for the binding of the ρ protein (fig. 12.8). Rho protein functions as a helicase, an enzyme that can separate RNA-DNA hybrid regions. After the *rut* site is synthesized in the RNA, ρ binds to the RNA and moves in the direction of RNA polymerase. In the DNA, a sequence, slightly upstream from the terminator, encodes an RNA sequence that forms a stem-loop structure containing several GC base pairs. As you may recall from chapter 9, a stem-loop structure can form due to complementary sequences within the RNA. These RNA stem-loops can form almost immediately after they are synthesized. The stem-loop causes RNA polymerase to pause in its synthesis of RNA. This allows ρ to catch up to the stem-loop, pass through it, and break the hydrogen bonds between the DNA and RNA

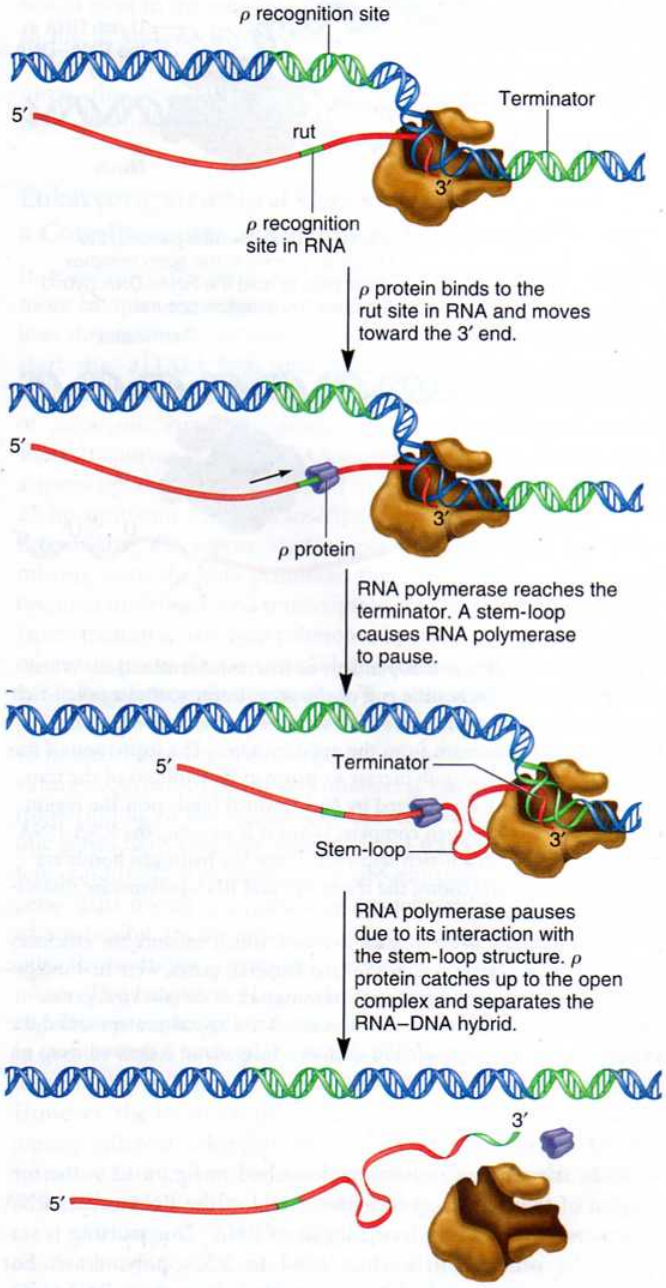


FIGURE 12.8 ρ-dependent termination.

within the open complex. When this occurs, the completed RNA strand is separated from the DNA along with RNA polymerase.

We now turn our attention to ρ-independent termination, a process that is facilitated by two sequence elements within the RNA (fig. 12.9). One element is a uracil-rich sequence located at the 3′ end of the RNA. The second sequence is slightly upstream from the uracil-rich sequence, near the 3′ end; it promotes the formation of a stem-loop structure.

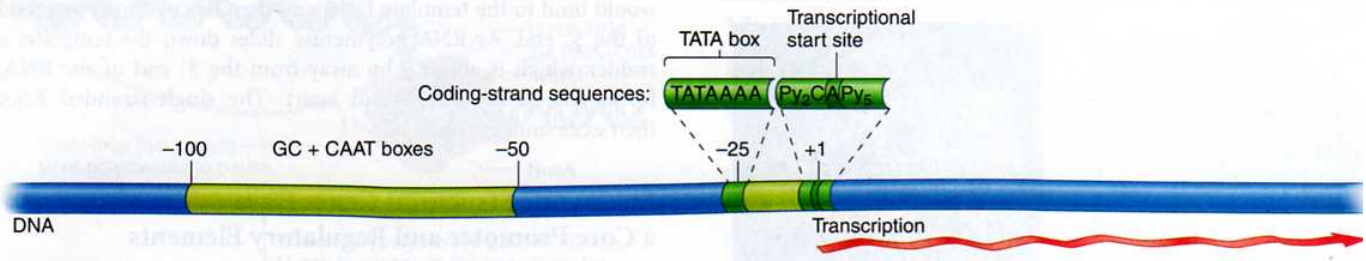


FIGURE 12.11 A common pattern found within the promoter of structural genes recognized by RNA polymerase II. The startpoint usually occurs at adenine; there are two pyrimidines and a cytosine that precede this adenine, and five pyrimidines that follow it. A TATA box is approximately 25 bp upstream. Regulatory elements, such as GC or CAAT boxes, are variable in their locations but often are found in the −50 to −100 region.

Transcription of Eukaryotic Structural Genes Is Initiated When RNA Polymerase II and General Transcription Factors Bind to a Promoter Sequence

Thus far, we have considered the DNA sequences that play a role in the promoter region of eukaryotic structural genes. By studying transcription in a variety of eukaryotic species, researchers have discovered that three categories of proteins are needed for basal transcription at the core promoter (table 12.2). These are RNA polymerase II, five different proteins called **general transcription factors (GTFs),** and a protein complex called **mediator.** We will examine their roles next.

Figure 12.12 describes the assembly of general transcription factors and RNA polymerase II at the TATA box. As shown here, a series of interactions leads to the formation of the open complex. Transcription factor IID (TFIID) first binds to the TATA box and thereby plays a critical role in the recognition of the promoter. TFIID is composed of several subunits including TATA-binding protein (TBP), which directly binds to the TATA box, and several other proteins called TBP-associated factors (TAFs). After TFIID binds to the TATA box, it associates with TFIIB. TFIIB promotes the binding of RNA polymerase II and TFIIF to the core promoter. Lastly, TFIIE and TFIIH bind to the complex. This completes the assembly of proteins to form a closed complex, also known as a **preinitiation complex.**

TFIIH plays a major role in the formation of the open complex. TFIIH has several subunits that perform different functions. One subunit hydrolyzes ATP and phosphorylates a domain in RNA polymerase II known as the carboxyl terminal domain (CTD). Phosphorylation of the CTD releases the contact between RNA polymerase II and TFIIB. Other subunits in TFIIH function as helicases, which break the hydrogen bonding between the double-stranded DNA and thereby promote the formation of the open complex. After the open complex has formed, TFIIB, TFIIE, and TFIIH dissociate. RNA polymerase II is then free to proceed to the elongation stage of transcription.

In vitro, when researchers mix together TFIID, TFIIB, TFIIE, TFIIF, TFIIH, RNA polymerase II, and a DNA sequence

TABLE 12.2

Proteins Needed for Transcription via the Core Promoter of Eukaryotic Structural Genes

RNA polymerase II: The enzyme that catalyzes the linkage of ribonucleotides in the 5′ to 3′ direction, using DNA as a template. Essentially all eukaryotic RNA polymerase II proteins are composed of 12 subunits. The two largest subunits are structurally similar to the β and β' subunits found in *E. coli* RNA polymerase.

General transcription factors:

TFIID: Composed of TATA-binding protein (TBP) and other TBP-associated factors (TAFs). Recognizes the TATA-binding sequence of eukaryotic structural gene promoters.

TFIIB: Binds to TFIID and then enables RNA polymerase II to bind to the core promoter.

TFIIF: Binds to RNA polymerase II and plays a role in its ability to bind to TFIIB and the core promoter. Also plays a role in the ability of TFIIE and TFIIH to bind to RNA polymerase II.

TFIIE: Plays a role in the formation and/or the maintenance of the open complex. It may exert its effects by facilitating the binding of TFIIH to RNA polymerase II and regulating the activity of TFIIH.

TFIIH: A multisubunit protein that has multiple roles. First, certain subunits have helicase activity and promote the formation of the open complex. Other subunits phosphorylate the CTD of RNA polymerase II, which releases its interaction with TFIIB and thereby allows RNA polymerase II to proceed to the elongation phase.

Mediator: A multisubunit complex that mediates the effects of regulatory transcription factors on the function of RNA polymerase II. Though mediator typically has certain core subunits, many of its subunits are variable. This variation may depend on the cell type and environmental conditions. The ability of mediator to affect RNA polymerase II function is thought to occur via the CTD of RNA polymerase II. Mediator can influence the ability of TFIIH to phosphorylate CTD, and subunits within mediator itself have the ability to phosphorylate CTD. Since CTD phosphorylation is needed to release RNA polymerase II from TFIIB, mediator plays a key role in the ability of RNA polymerase II to switch from the initiation to the elongation stage of transcription.

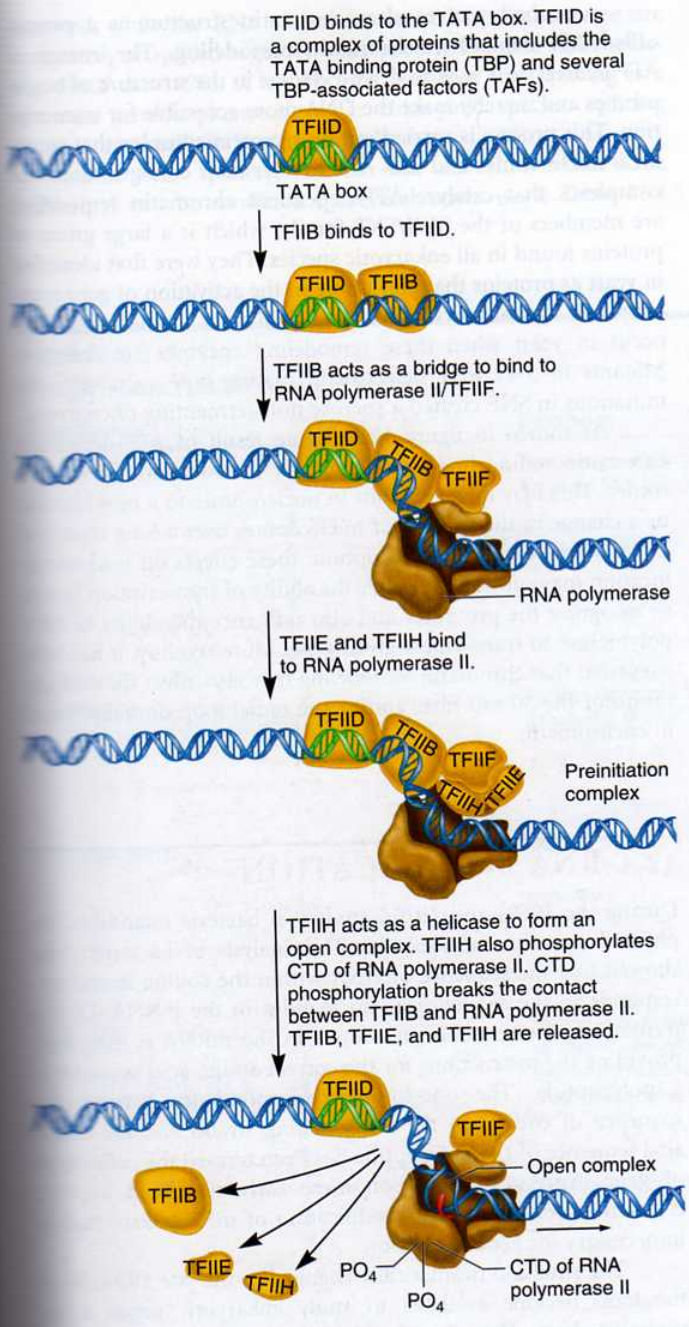


FIGURE 12.12 Steps leading to the formation of the open complex. TFIID first binds to the TATA box. A subunit of TFIID, known as the TATA-binding protein, recognizes the TATA box sequence. TFIIB then binds to TFIID. TFIIB promotes the binding of RNA polymerase II/TFIIF. Transcription factors TFIIE and TFIIH become bound to RNA polymerase to form the closed complex. To form the open complex, TFIIH hydrolyzes ATP and phosphorylates a region in RNA polymerase II known as the carboxyl terminal domain (CTD). RNA polymerase II is released from TFIIB. TFIIH also functions as a helicase, which breaks the hydrogen bonding between the double-stranded DNA and thereby promotes the formation of the open complex. After the open complex has formed, TFIIB, TFIIE, and TFIIH dissociate, and RNA polymerase II proceeds to the elongation stage of transcription.

containing a TATA box and transcriptional start site, the DNA is transcribed into RNA. Therefore, these components are referred to as the **basal transcription apparatus.** In a living cell, however, additional components regulate transcription and allow it to proceed at a reasonable rate.

A third component for transcription is a large protein complex termed *mediator.* The complex derives its name from the observation that it mediates interactions between RNA polymerase II and regulatory transcription factors that bind to enhancers or silencers. It serves as an interface between RNA polymerase II and many, diverse regulatory signals. The subunit composition of mediator is quite complex and variable (see table 12.2). The core subunits form an elliptical-shaped complex that partially wraps around RNA polymerase II. Mediator appears to regulate the ability of TFIIH to phosphorylate the CTD portion of RNA polymerase II. Therefore, it can play a pivotal role in the switch between transcriptional initiation and elongation. The function of mediator during eukaryotic gene regulation is explored in greater detail in chapter 15.

Chromatin Structure Plays a Key Role in Gene Transcription

As we have learned, transcription involves the binding of general transcription factors and RNA polymerase to the promoter region and the subsequent movement of RNA polymerase along the DNA double helix, allowing one strand to function as a template for transcription. The compaction of DNA to form chromatin, described in chapter 10, can be an obstacle to the transcription process. During interphase, when most transcription occurs, the chromatin of eukaryotes is found in 30 nm fibers that are organized into radial loop domains. Within the 30 nm fiber, the DNA is wound around histone octamers to form nucleosomes. The size of a histone octamer is roughly five times smaller than the complex of RNA polymerase II and GTFs. Therefore, since RNA polymerase is a very large enzyme compared to a nucleosome, the tight wrapping of DNA within a nucleosome is expected to inhibit the ability of RNA polymerase to transcribe the DNA. To circumvent this problem, the chromatin structure is significantly loosened during transcription.

Two common mechanisms alter chromatin structure (fig. 12.13). First, the amino terminal ends of histone proteins are covalently modified in a variety of ways including acetylation of lysines, methylation of lysines, and phosphorylation of serines. These covalent modifications play a key role in the compaction of chromatin. For example, positively charged lysine residues within the core histone proteins can be acetylated by enzymes called **histone acetyltransferases.** The attachment of the acetyl group ($-COCH_3$) may have two effects. First, it eliminates the positive charge on the lysine side chain and thereby disrupts the favorable interaction between the histone protein and the negatively charged DNA backbone. This may loosen the structure of nucleosomes and the 30 nm fiber and thereby facilitate the ability of RNA polymerase to transcribe a

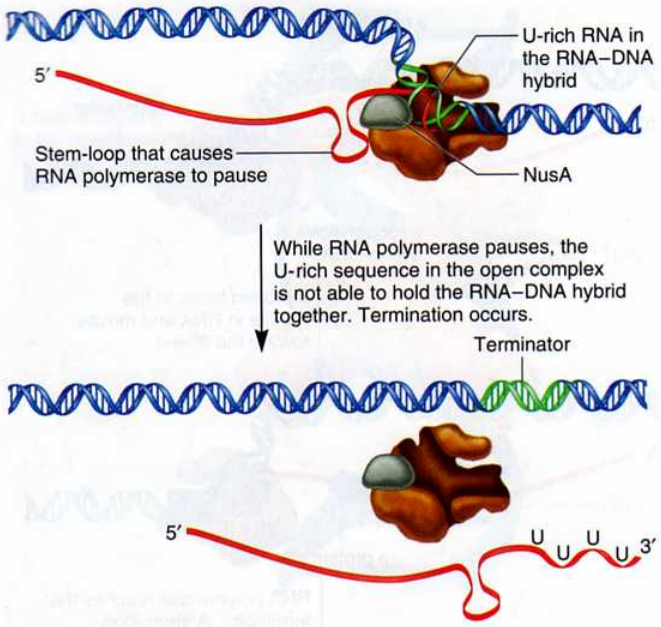


FIGURE 12.9 ρ-independent or intrinsic termination. When RNA polymerase reaches the end of the gene, it transcribes a uracil-rich sequence. Soon after this uracil-rich sequence is transcribed, a stem-loop forms just upstream from the open complex. The formation of this stem-loop causes RNA polymerase to pause in its synthesis of the transcript. This pausing is stabilized by NusA, which binds near the region where RNA exits the open complex. While it is pausing, the RNA-DNA hybrid region is a uracil-rich sequence. Since UA hydrogen bonds are relatively weak interactions, the transcript and RNA polymerase dissociate from the DNA.

Interestingly, proteins such as NusA, which enhance the efficiency of transcriptional termination in many bacterial genes, were first discovered because they actually prevent termination in certain viral genes. NusA aids in the prevention of termination via a viral protein called the N protein, which is discussed in chapter 14. Its name is derived from an acronym for N utilization substance.

In the sequence of events described in figure 12.9, the formation of the stem-loop near the 3′ end of the RNA causes RNA polymerase to pause in its synthesis of RNA. This pausing is stabilized by other proteins that bind to RNA polymerase. For example, a protein called NusA, which is bound to RNA polymerase, promotes pausing at stem-loop sequences. At the precise time RNA polymerase pauses, the uracil-rich sequence in the RNA transcript happens to be bound to the DNA template strand. As previously mentioned, the hydrogen bonding of RNA to DNA keeps RNA polymerase clamped onto the DNA. However, the binding of this uracil-rich sequence to the DNA template strand is thought to be unstable, causing the RNA transcript to spontaneously dissociate from the DNA and terminate further transcription. Because this process does not require a protein to physically remove the RNA transcript from the DNA, the sequence elements that cause this type of termination are referred to as **intrinsic terminators.**

12.3 TRANSCRIPTION IN EUKARYOTES

Many of the basic features of gene transcription are very similar in bacterial and eukaryotic species. Much of our understanding of transcription has come from studies in *Saccharomyces cerevisiae* (baker's yeast) and higher eukaryotic species such as mammals. In general, gene transcription in eukaryotes is more complex than that of their bacterial counterparts. Eukaryotic cells are larger and contain a variety of compartments known as organelles. This added level of cellular complexity dictates that eukaryotes contain many more genes encoding cellular proteins. In addition, higher eukaryotic species are multicellular, being composed of many different cell types. Multicellularity adds the requirement that genes be transcribed in the correct type of cell and during the proper stage of development. Therefore, in any given species, the transcription of the thousands of different genes that an organism possesses requires appropriate timing and coordination. In this section, we will examine features of gene transcription that are unique to eukaryotes. In addition, the regulation of eukaryotic gene transcription is covered in chapter 15.

Eukaryotes Have Multiple RNA Polymerases That Are Structurally Similar to the Bacterial Enzyme

The genetic material within the nucleus of a eukaryotic cell is transcribed by three different RNA polymerase enzymes, designated I, II, and III. Each of the three RNA polymerases transcribes different categories of genes. RNA polymerase I transcribes all of the genes that encode ribosomal RNA (rRNA) except for the 5S rRNA. RNA polymerase II plays a major role in cellular transcription since it transcribes all of the structural genes. It also transcribes certain snRNA genes, which are needed for pre-mRNA splicing. RNA polymerase II is responsible for the synthesis of all mRNA. RNA polymerase III transcribes all of the tRNA genes and the 5S rRNA gene.

All three RNA polymerases are very similar structurally and are composed of many subunits. They contain two large subunits that are similar to the β and β' subunits of bacterial RNA polymerase. The structures of a few RNA polymerases have been determined by X-ray crystallography. There is a remarkable similarity between the bacterial enzyme and its eukaryotic counterparts. Figure 12.10*a* compares the structures of a bacterial RNA polymerase with RNA polymerase II from yeast. As seen here, both enzymes have a very similar structure. Also, it is very exciting that this structure provides a way to envision how the transcription process works. As seen in figure 12.10*b*, the DNA that is about to be transcribed lies on a surface within RNA polymerase termed the bridge, which is nested between two regions called the jaw and clamp. A location called the wall forces the DNA-RNA hybrid to make a right-angle turn. This bend facilitates the ability of nucleotides to bind to the template strand. Mg^{++} is located at the catalytic site, which is precisely at the 3′ end of the growing RNA strand. Nucleoside triphosphates (NTPs) can enter the catalytic site via a funnel and pore region. The correct nucleotide

Figure 5.4

Action of *E. coli* RNA polymerase in the initiation and elongation stages of transcription.