CHAPTER SIX

Transcription of Genes

Genes are Expressed by Making RNA

Short Segments of the Chromosome Are Turned into Messages

Terminology: Cistrons, Coding Sequences and Open Reading Frames

How Is the Beginning of a Gene Recognized?

Manufacturing the Message

RNA Polymerase Knows Where to Stop

How Does the Cell Know Which Genes to Turn On?

What Activates the Activator?

Negative Regulation Results from the Action of Repressors

Many Regulator Proteins Bind Small Molecules and Change Shape

Transcription in Eukaryotes Is More Complex

Transcription of rRNA and tRNA in Eukaryotes

Transcription of Protein-Encoding Genes in Eukaryotes

Upstream Elements Increase the Efficiency of RNA Polymerase II Binding

Enhancers Control Transcription at a Distance

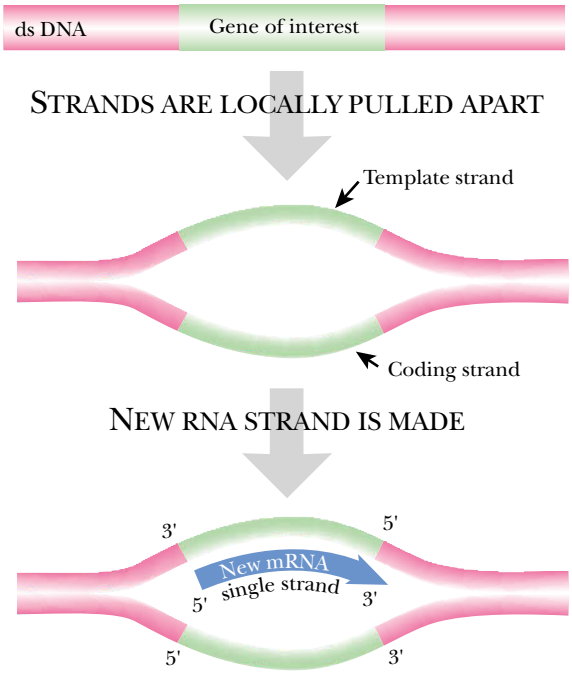


FIGURE 6.01 ***Transcription in Its Simplest Form***

The two strands of the DNA to be transcribed are separated locally. The top strand serves as a template for building a new RNA molecule.

Genes are Expressed by Making RNA

DNA "merely" stores genetic information. Putting the information to use requires RNA and (usually) protein.

For a cell to operate, its genes must be expressed. The word "expressed" means that the gene products, whether proteins or RNA molecules, must be made. The DNA molecule that carries the original copy of the genetic information is used to store genetic information but is not used as a direct source of instructions to run the cell. Instead, working copies of the genes, made of RNA, are used. The transfer of information from DNA to RNA is known as **transcription** and RNA molecules are therefore sometimes referred to as transcripts. Genes may be subdivided into two major groups: those whose final product is an RNA molecule (e.g., tRNA, rRNA, assorted regulatory RNAs—see below) and those whose final product is protein. In the latter case, the RNA transcript acts as an intermediary and a further step is needed to convert the information carried by the RNA to protein. This process is discussed in Ch. 8. The type of RNA molecule that carries genetic information encoding a protein from the genes into the rest of the cell is known as **messenger RNA**, or mRNA. Since the great majority of genes encode proteins, we will deal with these genes first.

Messenger RNA carries the information for making proteins from the genes to the cytoplasm.

For a gene to be transcribed, the DNA, which is double stranded, must first be pulled apart temporarily, as shown in Figure 6.01. Then, RNA is made by **RNA polymerase**. This enzyme binds to the DNA at the start of a gene and opens the double helix. Finally, it manufactures an RNA molecule.

The DNA double helix must be opened up for RNA polymerase to read the template strand and make RNA.

The sequence of the RNA message is complementary to the **template strand** of the DNA from which it is synthesized. Apart from the replacement of thymine in DNA with uracil in RNA, this means that the sequence of the new RNA molecule is identical to the sequence of the **coding strand** of DNA; that is, the strand not actually used as a template during transcription. Note that RNA, like DNA, is synthesized in the 5′- to 3′- direction (Fig. 6.02). Other names for the template strand are the *non-coding* or *anti-sense* strand; other names for the coding strand are *non-template* or *sense* strand. Only one of the strands of DNA is copied in any given transcribed region. [But note

coding strand The strand of DNA equivalent in sequence to the messenger RNA (same as plus strand)
messenger RNA (mRNA) The molecule that carries genetic information from the genes to the rest of the cell
RNA polymerase Enzyme that synthesizes RNA using a DNA template
transcription Process by which information from DNA is converted into its RNA equivalent
template strand Strand of DNA used as a guide for synthesizing a new strand by complementary base pairing

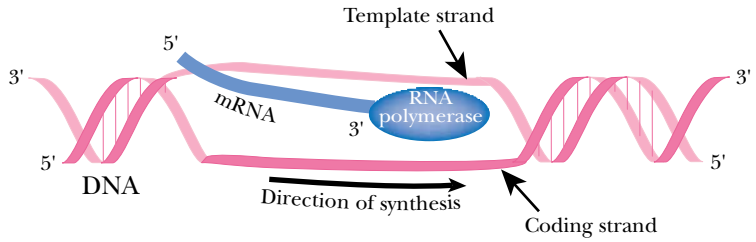


FIGURE 6.02 ***Naming the Basic Components Involved in Transcription***

The DNA is shown in its double helical form. After local separation of the strands, the new RNA is synthesized so that it base pairs with one of the DNA strands—the template strand. The other DNA strand is inactive and is called the coding strand. The enzyme RNA polymerase synthesizes single-stranded RNA in the 5′- to 3′-direction. The sequence of bases in the RNA is the same as in the coding strand of DNA, except that uracil substitutes in RNA for thymine in DNA. The bases of the mRNA are complementary to those of the template DNA strand; note that uracil base pairs with adenine.

that the two different strands of the DNA may each be used as templates in different regions of the chromosome.]

Short Segments of the Chromosome Are Turned into Messages

Each mRNA carries information from only a short stretch of the DNA.

Although a chromosome carries hundreds or thousands of genes, only a fraction of these are in use at any given time. In a typical bacterial cell, about 1,000 genes, or about 25% of the total, are expressed under any particular set of growth conditions. Some genes are required for the fundamental operations of the cell and are therefore expressed under most conditions. These are known as **housekeeping genes**. Other genes vary in expression in response to changes in the environment. During cell growth and metabolism, each gene or small group of related genes is used to generate a separate RNA copy when, and if, it is needed. Consequently, each cell contains many different RNA molecules, each carrying the information from a short stretch of DNA.

In the cells of higher organisms, which have many more genes than do bacteria, the proportion of genes in use in a particular cell at a particular time is much smaller. Different cells of multi-cellular organisms express different selections of genes depending on their specialized roles. For example, in the human female, genes related to the menstrual cycle are largely unique and expressed in a timed sequence to provide functionality to the organism. In addition, gene expression varies with the stage of development. Embryonic genes are often expressed only at certain times. Thus, the control of gene expression is much more complex in higher organisms, although the basic principles are the same.

Terminology: Cistrons, Coding Sequences and Open Reading Frames

A **cistron** is a **structural gene**, which is a coding sequence or segment of DNA encoding a polypeptide. It was defined originally as a genetic unit by complementation using the *cis/trans* test. Nowadays, the terms cistron and structural gene also include DNA sequences that code for RNA molecules that function as RNA without being translated into proteins (e.g., rRNA, tRNA, snRNA, etc.). An **open reading frame** (**ORF**)

cistron Segment of DNA (or RNA) that encodes a single polypeptide chain
housekeeping genes Genes that are switched on all the time because they are needed for essential life functions
open reading frame (ORF) Sequence of bases (either in DNA or RNA) that can be translated (at least in theory) to give a protein
structural gene Sequence of DNA (or RNA) that codes for a protein or for an untranslated RNA molecule

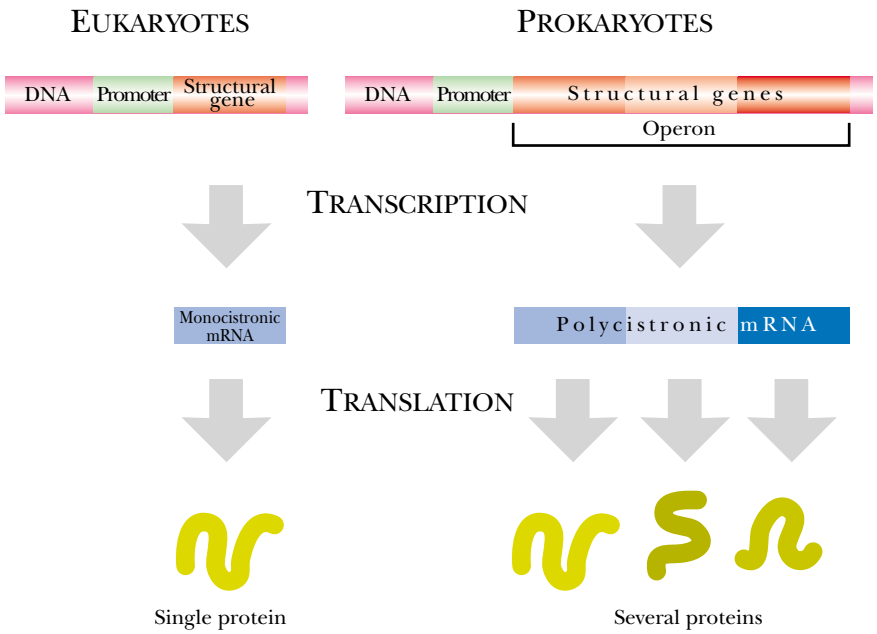


FIGURE 6.03
Monocistronic Versus Polycistronic mRNA

The typical situation in eukaryotes is to have one structural gene produce monocistronic RNA and this, in turn, be translated into a single protein. In bacteria, it is common to see several structural genes transcribed under the control of a single promoter. The RNA produced is polycistronic and yields several separate proteins.

is any sequence of bases (in DNA or RNA) that could, in theory, encode a protein. The ORF is "open" in the sense that it does not contain any stop codons that would interrupt its translation into a polypeptide chain (although, of course, every ORF ends in a stop codon). (Any cistron that encodes a protein must also be an ORF, whereas a cistron that encodes an untranslated RNA is not an ORF.)

In eukaryotes, each mRNA carries only a single gene. In prokaryotes, several genes may be carried on the same mRNA.

In eukaryotes, each gene is transcribed to give a separate mRNA and each mRNA molecule therefore encodes the information for only a single protein and is known as **monocistronic mRNA** (Fig. 6.03). In bacteria, clusters of related genes, known as **operons**, are often found next to each other on the chromosome and are transcribed together to give a single mRNA, which is therefore called **polycistronic mRNA**. Thus, a single bacterial mRNA molecule may encode several proteins, usually with related functions, such as the enzymes that oversee the successive steps in a metabolic pathway.

How Is the Beginning of a Gene Recognized?

Before starting transcription, RNA polymerase binds to the promoter, a recognition sequence in front of the gene.

Transcription will first be described in bacteria because it is simpler. The principles of transcription are similar in higher organisms, but the details are more complicated, as will be shown below. The major differences between prokaryotes and eukaryotes occur in the initiation and regulation of transcription, rather than in the actual synthesis of RNA. In front of each gene is a region of regulatory DNA that is not itself transcribed. This contains the **promoter**, the sequence to which RNA polymerase binds (Fig. 6.04), together with other sequences involved in the control of gene expression. This stretch of DNA in front of a gene (i.e., at the 5′-end) is often referred to as the **upstream region**. Note also that the first base of the mRNA of a protein-encoding gene is not the first base of the protein coding sequence. Between these two points there is a short stretch known as the **5′-untranslated region**, or 5′-UTR, meaning it will not be translated to form protein (Fig. 6.04). At the far end of the mRNA there is another short

5′-untranslated region (5′-UTR) Region of an mRNA between the 5′-end and the translation start site
monocistronic mRNA mRNA carrying the information of a single cistron, that is a coding sequence for only a single protein
operon A cluster of prokaryotic genes that are transcribed together to give a single mRNA (i.e. polycistronic mRNA)
polycistronic mRNA mRNA carrying the information of multiple cistrons, that is coding sequences for several proteins
promoter Region of DNA in front of a gene that binds RNA polymerase and so promotes gene expression
upstream region Region of DNA in front (i.e. beyond the 5′-end) of a structural gene; its bases are numbered negatively counting backwards from the start of transcription

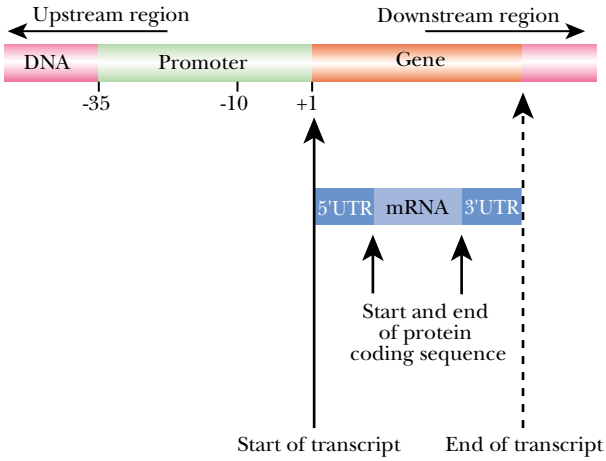


FIGURE 6.04 ***Upstream and Downstream Regions***

Genes and their regulatory regions are divided into upstream and downstream portions. The upstream portion contains the promoter. The downstream region begins with the information for the 5′-untranslated component, then the structural gene. The messenger RNA begins with the 5′-untranslated region (5′-UTR), then the coding sequence for the protein. Transcription begins by definition at the first base after the promoter. The upstream region, including the promoter, is given negative numbers counting backward from the beginning of transcription.

region, beyond the end of the protein coding sequence, that is not translated. This is the **3′-untranslated region**, or 3′-UTR.

The sigma subunit of bacterial RNA polymerase recognizes the promoter. The core enzyme makes RNA.

Bacterial RNA polymerase consists of two major components, the **core enzyme** (itself made of four subunits) and the **sigma subunit**. The core enzyme is responsible for RNA synthesis whereas the sigma subunit is largely responsible for recognizing the promoter. The sigma subunit, recognizes two special sequences of bases in the promoter region of the coding (non-template) strand of the DNA (Fig. 6.05). These are known as the **–10 sequence** and the **–35 sequence** because they are found by counting backward approximately 10 and 35 bases, respectively, from the first base that is transcribed into mRNA. [Previously, the –10 sequence was known as the **Pribnow box**, after its discoverer. This name is rarely used nowadays.]

Strong promoters usually have sequences close to consensus.

The consensus sequence for the –10 sequence is TATAA and the consensus sequence at –35 is TTGACA. (Consensus sequences are found by comparing many sequences and taking the average.) Although a few highly expressed genes do have the exact consensus sequences in their promoters, the **–10** and **–35** region sequences are rarely perfect. However, as long as they are wrong by only up to three or four bases, the sigma subunit will still recognize them. *The strength of a promoter depends partly on how closely it matches the ideal consensus sequence*. Strong promoters are highly expressed and are often close to consensus. Promoters further away from the consensus sequence will be expressed only weakly (in the absence of other factors—but see below).

Promoter sequences vary in different organisms.

In practice, consensus sequences for regulatory sites on DNA such as promoters will vary from one group of organisms to another. Thus, the –10 and –35 consensus sequences given above are for *Escherichia coli* and related bacteria. Both the consensus sequences and the proteins that recognize them will diverge in more distantly related organisms. This is of practical importance when genes from one organism are expressed in another as a result of biotechnological manipulations. Consequently, it is

–10 region Region of bacterial promoter 10 bases back from the start of transcription that is recognized by RNA polymerase
3′-untranslated region (3′-UTR) Sequence at the 3′-end of mRNA, downstream of the final stop codon, that is not translated into protein
–35 region Region of bacterial promoter 35 bases back from the start of transcription that is recognized by RNA polymerase
core enzyme Bacterial RNA polymerase without the sigma (recognition) subunit
Pribnow box Another name for the –10 region of the bacterial promoter
sigma subunit Subunit of bacterial RNA polymerase that recognizes and binds to the promoter sequence

FIGURE 6.05 ***Sigma Recognizes the –10 and –35 Sequences***

The sigma protein binds to both the –10 and –35 sequences of the promoter, thereby establishing a constant position with respect to the start of transcription.

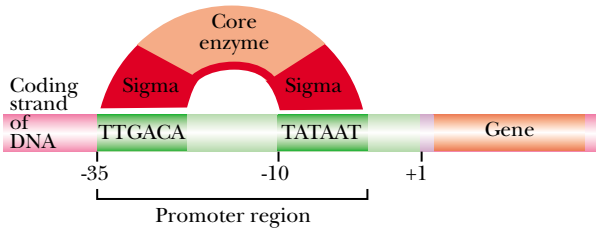
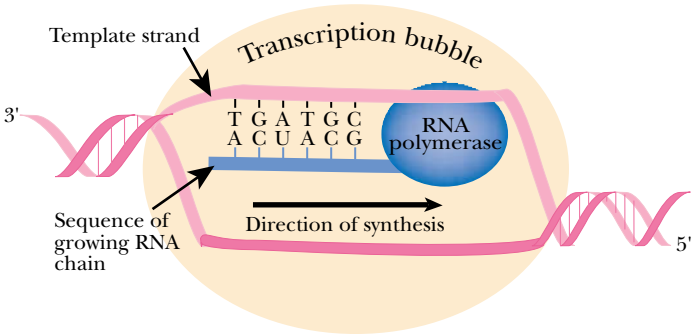


FIGURE 6.06 ***Elongation of the mRNA***

The beginning of RNA synthesis is shown. The DNA strands have separated at the transcription bubble. Synthesis of six bases of RNA complementary to those of the DNA template strand occurs while RNA polymerase remains at the promoter site.



often helpful to supply consensus promoters (or other regulatory sequences) that work well in the host organism to achieve high expression of a cloned gene from a foreign source. The use of "expression vectors" to optimize gene expression during cloning is discussed in more detail in Ch. 22.

Manufacturing the Message

RNA polymerase opens up the DNA to form the transcription bubble.

Once the sigma subunit has bound to a promoter, the RNA polymerase core enzyme opens up the DNA double helix locally to form the **transcription bubble**. Note that the –10 sequence, TATAAT, consists of AT base pairs and this assists in the melting of the DNA into single strands. After the DNA helix has been opened, a single strand of RNA is generated using one of the DNA strands as a template for matching up the bases. Once the RNA polymerase has bound to the DNA and initiated a new strand of RNA, the sigma subunit is no longer needed and often (though not always) detaches from the DNA, leaving behind the core enzyme. The RNA polymerase actually remains at the promoter until the new strand is eight or nine bases long. At this point, sigma leaves and the core enzyme is free to move forward and elongate the mRNA (Fig. 6.06).

The core enzyme moves ahead, manufacturing RNA and leaving sigma behind at the promoter.

The first transcribed base of the mRNA is normally an A (as in Fig. 6.06). This special A is usually flanked by two pyrimidines, most often giving the sequence CAT. Sometimes the first transcribed base is a G, but almost never a pyrimidine. Synthesis of mRNA is from 5′ to 3′ and proceeds at about 40 nucleotides per second. This is much slower than DNA replication (~1,000 bp/sec), but roughly equivalent to the rate of polypeptide synthesis (15 amino acids per second).

The core enzyme of RNA polymerase consists of four subunits, two α plus β and β′ (Fig. 6.07). The β and β′ subunits comprise the catalytic site of the enzyme. The α subunit is required partly for assembly and partly for recognizing promoters. RNA polymerase has a deep groove through the middle that can accommodate about 16 bp of DNA in the case of bacteria and about 25 bp in the case of eukaryotes such as yeast, whose RNA polymerase is larger. A thinner groove, roughly at right angles to the first, may hold the newly constructed strand of RNA.

transcription bubble Region where DNA double helix is temporarily opened up so allowing transcription to occur

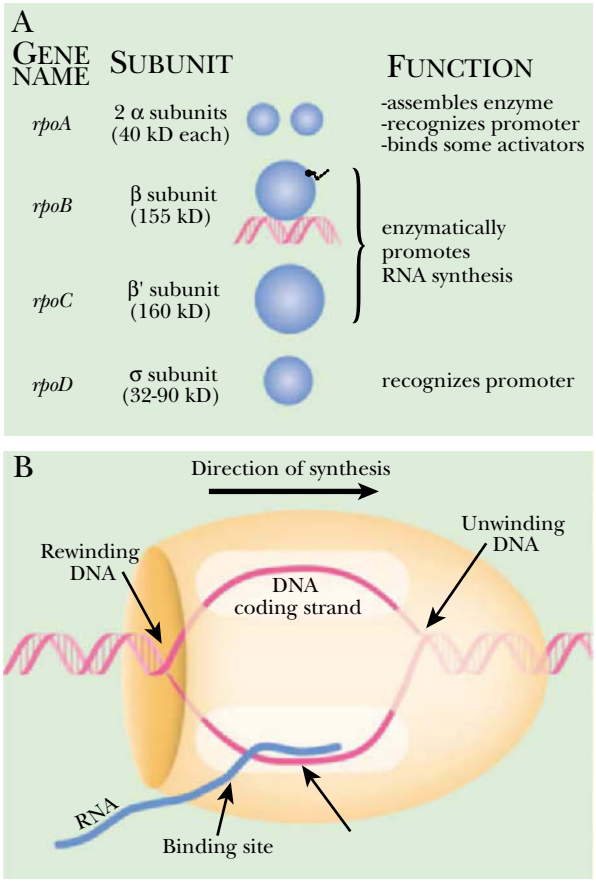


FIGURE 6.07 ***Structure of RNA Polymerase***

A) Bacterial RNA polymerase has four types of subunits and three functional specificities. B) Topography of functions performed by RNA polymerase. C) The structure of yeast RNA polymerase reveals a groove that may be used by DNA as it moves relative to the polymerase and a potential groove for the newly formed RNA.

The negative supercoiling of the chromosome promotes opening of DNA during transcription. As RNA polymerase moves along the DNA, it winds the DNA more tightly ahead of itself, creating positive supercoils. It also leaves partly unwound DNA behind, which generates negative supercoils. To restore normal levels of supercoiling, DNA gyrase inserts negative supercoils ahead of RNA polymerase and topoisomerase I removes negative supercoils behind RNA polymerase (see Ch. 4).

RNA Polymerase Knows Where to Stop

The end of a gene is marked by a terminator sequence that forms a hairpin structure in the RNA.

Just as there is a recognition site at the front of each gene, so there is a special **terminator** sequence at the end. The terminator is in the template strand of DNA and consists of two inverted repeats separated by half a dozen bases, and followed by a run of A's. The sequence of the mRNA will be the same as the non-template strand of DNA except for the substitution of U for T. Thus the string of A's in the DNA template strand gives rise to a run of U's at the 3′-end of the mRNA (Fig. 6.08). Note that in the DNA, the two inverted repeat sequences are on opposite strands. Although researchers often talk as if the mRNA has inverted "repeats," its second "repeat" is actually the complement of the first. Because of this, such inverted repeats on the

terminator DNA sequence at end of a gene that tells RNA polymerase to stop transcribing

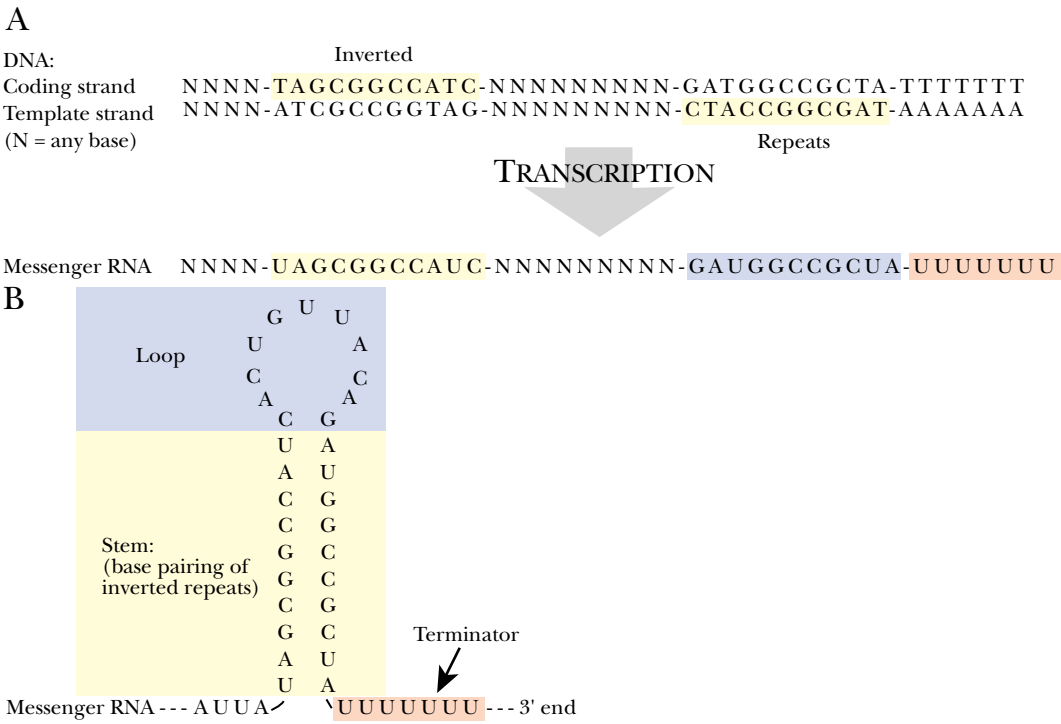


FIGURE 6.08 ***The Terminator Sequence is Transcribed into RNA***

A) The signal for RNA polymerase to stop is shown in both the DNA and the RNA transcribed from it. The terminator consists of an inverted repeat separated by approximately 10 bases from a run of U's. B) The complementary bases form the stem of the hairpin, with the intervening bases forming the loop.

same strand of an RNA molecule can pair up to generate a stem and loop or "hairpin" structure (Fig. 6.08).

Once the RNA polymerase reaches the stem and loop, it pauses. Long RNA molecules contain many possible hairpin structures that cause RNA polymerase to slow down or stop briefly, depending on the size of the hairpin. This provides an opportunity for termination, but if there is no string of U's, the RNA polymerase will start off again. However, a string of U's paired with a string of A's in the template strand of DNA is a very weak structure, and the RNA and DNA fall apart while the RNA polymerase is idling (Fig. 6.09). Pausing varies in length, but is around 60 seconds for a typical terminator.

Termination may actually occur at several possible positions in the middle or end of the run of U's. In other words, the RNA polymerase "stutters" and the precise location of termination may vary slightly between different molecules of the same mRNA. Once the DNA and RNA have separated at the terminator structure, the RNA polymerase falls off and departs to find another gene (Fig. 6.09).

Two classes of terminators exist. **Rho-dependent terminators** need **Rho (ρ) protein** to separate the RNA polymerase from the DNA. **Rho-independent**, or "intrinsic," terminators do not need Rho or any other factor to cause termination. Most terminators in *E. coli* do not need Rho. In contrast, Rho-dependent terminators are relatively frequent in bacteriophages.

A sub-class of terminators require a recognition protein, known as Rho, to function.

Rho protein is a specialized helicase that uses energy from ATP to unwind a DNA/RNA hybrid double helix. It consists of a hexamer of six identical subunits that recognizes and binds to a sequence of 50 to 90 bases located upstream of the termi-

Rho (ρ) protein Protein factor needed for successful termination at certain transcriptional terminators
Rho-dependent terminator Transcriptional terminator that depends on Rho protein
Rho-independent terminator Transcriptional terminator that does not need Rho protein

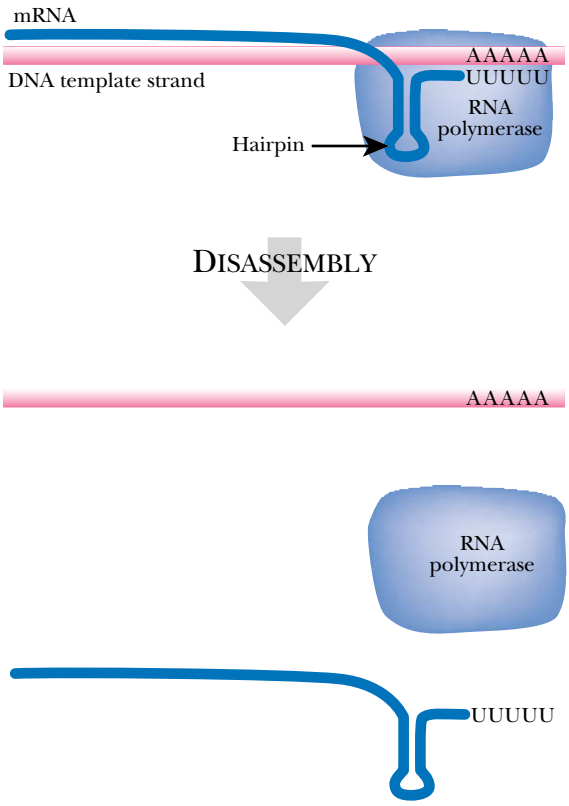


FIGURE 6.09 ***Termination of Messenger RNA***

When the mRNA reaches the hairpin of the terminator it pauses; when it reaches the AAAAA sequence it falls off the template strand along with the newly synthesized RNA.

nator in the mRNA. The Rho hexamer does not form a closed ring, but instead is split open and resembles a lock washer in structure. The RNA sequence for Rho binding is poorly defined but is high in C and low in G. Rho can only bind to the growing mRNA chain once the RNA polymerase has synthesized the C-rich/G-poor recognition region and moved on. Rho moves along the RNA transcript and catches up with the RNA polymerase at the terminator stem and loop structure where the RNA polymerase pauses (Fig. 6.10). Rho then unwinds the DNA/RNA helix in the transcription bubble and separates the two strands.

How Does the Cell Know Which Genes to Turn On?

Housekeeping genes are switched on all the time.

Some genes, known as housekeeping genes, are switched on all the time; i.e., they are expressed "**constitutively**." In bacteria, these often have both their –10 and –35 region promoter sequences very close or identical to consensus. Consequently, they are always recognized by the sigma subunit of RNA polymerase and are expressed under all conditions. Other constitutive promoters are further from consensus and expressed less strongly. Nonetheless, if only relatively low amounts of the gene product are needed, this is acceptable.

Some genes need activator proteins to switch them on.

Genes that are only needed under certain conditions sometimes have poor recognition sequences in the –10 and –35 regions of their promoters. In such cases, the promoter is not recognized by sigma unless another accessory protein is there to help (Fig. 6.11). These accessory proteins are known as gene **activator proteins** and are different for different genes. Each activator protein may stimulate the transcription of one or more genes. A group of genes that are all recognized by the same activator protein will be expressed together under similar conditions, even if the genes are at different places on the DNA. Higher organisms have many genes that are expressed differently

activator protein Protein that switches a gene on
constitutive gene Gene that is expressed all the time

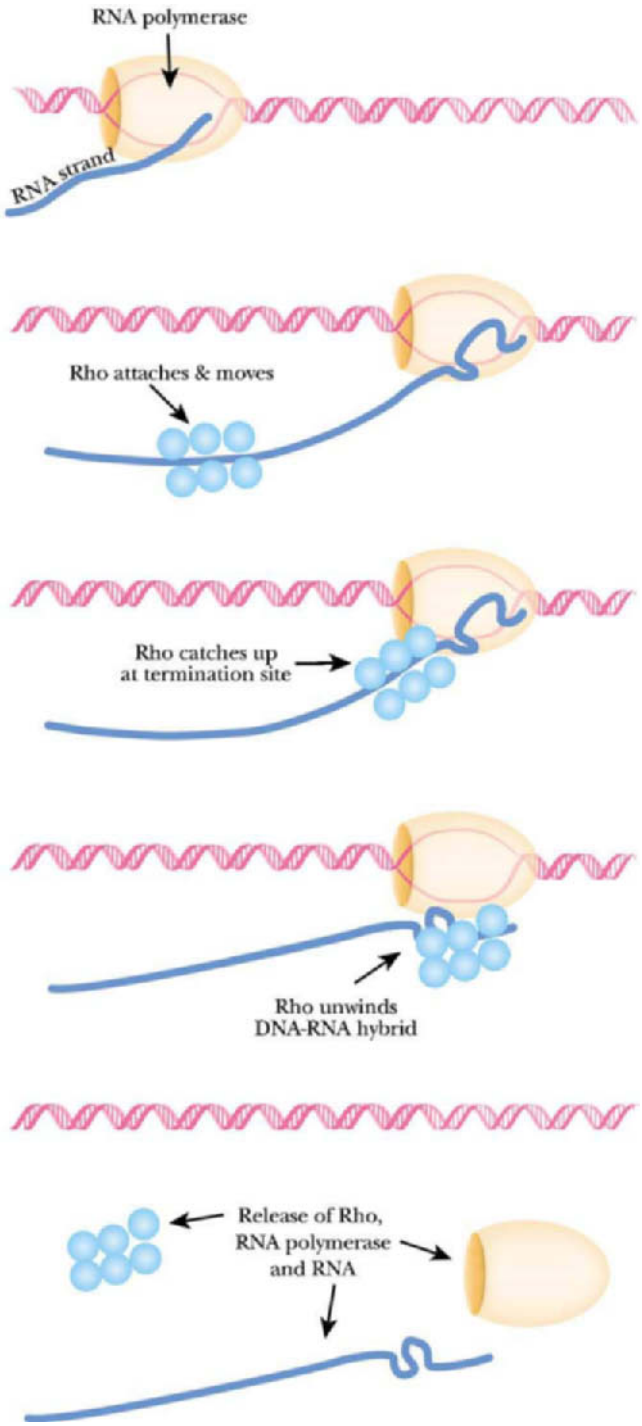


FIGURE 6.10 ***Termination by Rho***

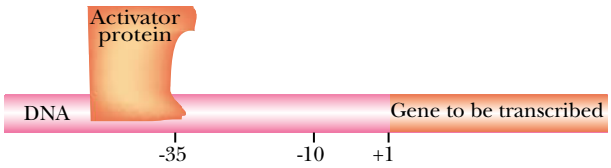
Rho first binds to the growing messenger RNA. When the RNA polymerase pauses at the termination site, Rho catches up and untwists the newly formed mRNA strand from the DNA. Subsequently, the mRNA and RNA polymerase fall off the DNA and Rho detaches from the mRNA.

in different tissues. As a result, eukaryotic genes are often controlled by multiple activator proteins, more specifically known as **transcription factors** (see below).

What Activates the Activator?

Long ago, the Greek philosopher Plato pondered the political version of this question: "Who will guard the guardians?" In living cells, especially in more complex higher

transcription factor Protein that regulates gene expression by binding to DNA in the control region of the gene



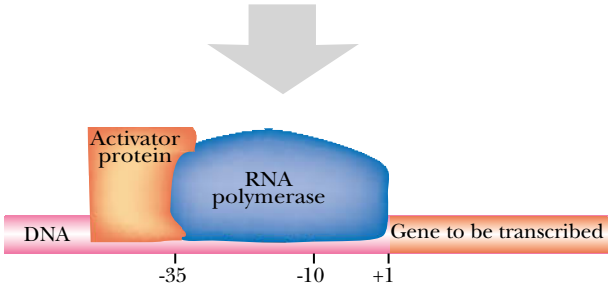


FIGURE 6.11 ***Gene Activator Proteins***

The activator protein first binds to the promoter region of the gene. Once bound, the activator protein facilitates the binding of the RNA polymerase. Gene transcription then commences.

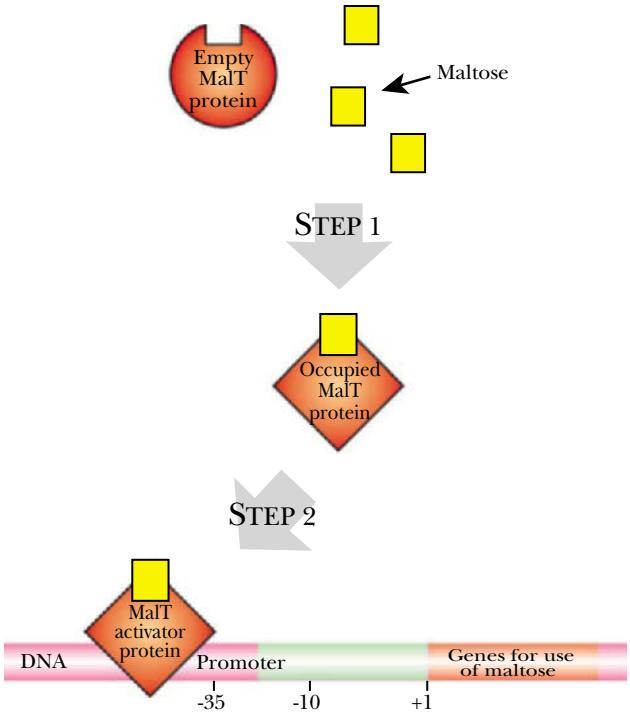


FIGURE 6.12 ***MalT Changes Shape upon Binding Maltose***

The MalT protein has a binding site complementary in shape to the sugar maltose. In step 1, MalT binds to maltose, which causes MalT to change shape. In step 2, the new conformation of MalT protein allows it to bind to DNA at a specific sequence found only in certain promoters. The gene thus activated is involved in the metabolism of maltose.

organisms, there may indeed be a series of regulators, each regulating the next. What is the initial event? The cell must respond to some outside influence or must be influenced by other internal processes. The regulation of gene expression will be considered in more detail in Chapters 9 and 10. This chapter will be limited to a discussion of the basic mechanisms needed for a promoter to be functional.

As a simple example of an activator, consider the use of maltose by *Escherichia coli*. Maltose is a sugar made originally from the starch in malt and many other sources. It can be used by *E. coli* to satisfy all of its needs for energy and organic material. An activator protein, MalT, detects maltose and binds to it (Fig. 6.12). This causes the MalT protein to change shape, exposing its DNA-binding site. The original "empty" form of MalT cannot bind to DNA. The active form (MalT + maltose) binds to a specific sequence of DNA found only in the promoter region of genes needed for growth on maltose. The presence of MalT helps RNA polymerase bind to the promoter and transcribe the genes. The small molecule, in this case maltose, which causes gene

Activator proteins often change shape in response to small molecules. Only one conformation binds to DNA.

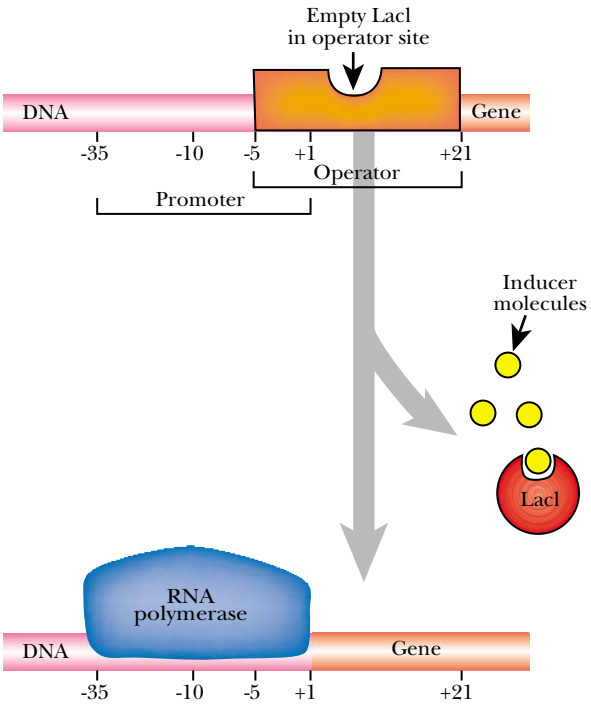


FIGURE 6.13 ***Principle of Negative Regulation by a Repressor***

The LacI protein is bound to the operator site, within the promoter region of a gene that affects lactose metabolism. The inducer binds to LacI, changing its conformation and causing its release from the DNA. The RNA polymerase is then free to transcribe the gene.

expression is known as the **inducer**. The result of this is that the genes intended for using maltose are only induced when this particular sugar is available. The same general principle applies to most nutrients, although the details of the regulation may vary from case to case.

Negative Regulation Results from the Action of Repressors

Repressors are proteins that switch genes off.

Genes may be controlled by positive or **negative regulation**. In **positive regulation**, an activator protein binds to the DNA only when the gene is to be turned on. In **negative regulation**, a **repressor** protein binds to the DNA and insures that the gene is turned off. Only when the repressor is removed from the DNA can the gene be transcribed. The site where a repressor binds is called the **operator** sequence. Like activator proteins, repressor proteins alternate between DNA-binding and nonbinding forms. In this case, binding of the inducer to the repressor causes it to change from its DNA-binding form to the nonbinding form.

Historically, negative regulators were discovered before activators. The best known example is the lactose repressor, the **LacI protein** (Fig. 6.13). Lactose is another sugar, found in milk, on which bacteria such as *E. coli* can grow. When no lactose is available, the LacI protein binds to its operator sequence, which overlaps part of the promoter and the front part of the coding region for the genes for using lactose. When lactose is present, the LacI protein changes shape and is released from the DNA and the lactose genes are induced. Overall, the result is the same as for maltose: when lactose is available, the genes for using it are switched on and when there is no lactose, the genes are turned off.

inducer Small signal molecule that binds to a regulatory protein and thereby causes a gene to be switched on
LacI protein Repressor that controls the *lac* operon
negative regulation Regulatory mode in which a repressor keeps a gene switched off until it is removed
operator Site on DNA to which a repressor protein binds
positive regulation Control by an activator that promotes gene expression when it binds
repressor Regulatory protein that prevents a gene from being transcribed

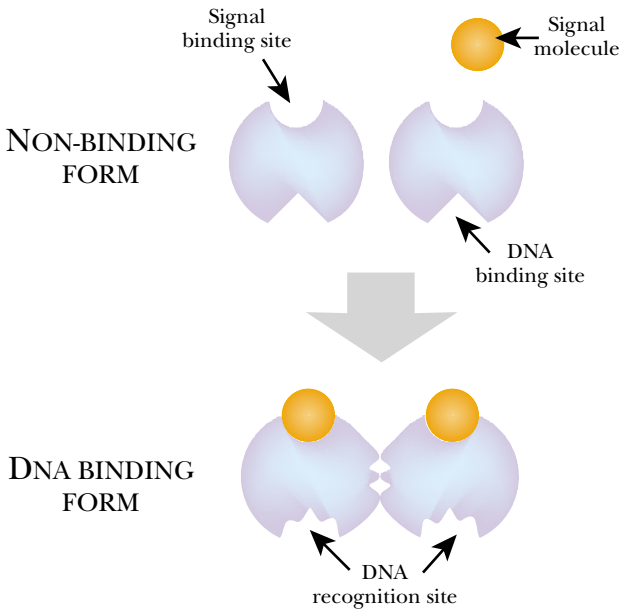


FIGURE 6.14 ***Allosteric Protein Binds a Signal Molecule and Changes Shape***

The two subunits shown have a signal-binding site and a DNA-binding site. When the signal molecule binds to the subunits, they pair and change conformation They are then able to bind to DNA.

The detailed mechanism by which repressors prevent transcription varies considerably and is often unknown. The repressor sometimes blocks the binding of RNA polymerase to the promoter, simply by getting in the way (steric hindrance). An example is the well-studied CI repressor of bacteriophage lambda. Sometimes the repressor may bind further downstream, inside the structural gene. In this case, RNA polymerase can still bind to the promoter but is prevented from moving forward and transcribing the gene. Sometimes, even though their binding sites in the DNA sequence overlap, the RNA polymerase and the repressor both bind the DNA simultaneously. [Remember that the DNA double helix is 3-dimensional and that two proteins may therefore bind to the same linear segment, if they occupy separate locations around its surface.] Indeed, an example of this is the LacI repressor. In this case the RNA polymerase actually binds more tightly in the presence of repressor, but is locked in place and cannot open the DNA to initiate transcription.

Many Regulator Proteins Bind Small Molecules and Change Shape

Small molecules may control gene expression by binding to regulatory proteins.

Whether a regulator protein is an activator or a repressor, it needs a signal of some sort. One of the most common ways to do this is by using some small molecule that fits into a binding site on the regulatory protein (Fig. 6.14). This is called the **signal molecule**. In the case of using a nutrient for growth, an obvious and common choice is the nutrient molecule itself. [In prokaryotes the DNA binding protein often binds the signal molecule directly. In eukaryotes, where the DNA is inside the nucleus, things are often more complex, and multiple proteins are involved. The signal molecule is often bound by proteins in the cell membrane or cytoplasm and the signal is then transmitted to the nucleus. The DNA binding protein itself normally stays in the nucleus and upon receiving the signal, is converted to its DNA-binding form by phosphorylation.]

When a regulator protein binds its signal molecule, it changes shape (Fig. 6.14). Regulator proteins have two alternative forms, the DNA-binding form and the non-binding form. Binding or loss of the signal molecule causes the larger protein to flip-flop between its two alternative shapes. Proteins that change in activity by changing

signal molecule Small molecule that exerts a regulatory effect by binding to a regulatory protein

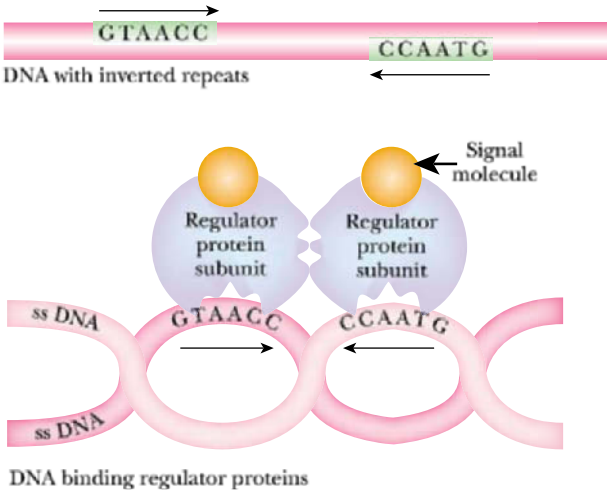


FIGURE 6.15 ***Regulator Binds at an Inverted Repeat—Principle***

At sites where regulator proteins bind there is often an inverted repeat with both DNA strands participating as shown. If the subunits of regulator protein are identical, they each recognize one of the inverted repeats and pair so that the same regions of each subunit face each other.

shape in this manner are called **allosteric proteins**. Examples include some enzymes, transport proteins and regulators. Allosteric proteins have multiple subunits that change shape in concert (Fig. 6.14). Usually there is an even number of subunits, most often two or four. All of the subunits bind the signal molecule and then they all change shape together.

Recognition sites on DNA are often inverted repeats. Separate subunits of the regulator protein each bind one of the repeat sequences.

Since there is an even number of protein subunits, the recognition site on the DNA for regulator proteins is often duplicated. In this case, the recognition site is usually an inverted repeat, often referred to as a palindrome. This is because the subunits of the regulator protein bind to each other head to head rather than head to tail (Fig. 6.15). Consequently, the two protein molecules are pointing in opposite directions. Because they have identical binding sites for DNA, they recognize the same sequence of bases but in opposite directions on the two strands of the DNA. The two half-sites are usually separated by a spacer region of several bases, whose identity is free to vary. The two half-sequences of such recognition sites are not always exact matches.

One regulator protein subunit binds to the recognition sequence on the template strand of the DNA double helix, and its partner binds to the same sequence but on the non-template strand of the DNA pointing in the opposite direction. This is simpler in practice than it sounds, precisely because the DNA molecule is helical. Although the two recognition sequences are on different strands of DNA, they end up on the same face of the DNA molecule due to its helical twisting (Fig. 6.15).

An example of a palindromic recognition site is the *lac* operator sequence which is bound by the LacI repressor (a tetramer). This sequence runs from –6 to +28 relative to the start of transcription. It is not exactly symmetrical. The two half sequences are: TGTGTGgAATTGTgA and, running in the opposite sense on the other strand, TGTGTGaAATTGTtA (capital letters indicate matching bases). The two half-sites are separated by five base pairs. The left hand half of this site binds the LacI protein more strongly than the right hand side. A stronger operator sequence could be generated by changing the right hand half-site to exactly match the left.

Transcription in Eukaryotes Is More Complex

Eukaryotes have three RNA polymerases that specialize in which type of genes they transcribe.

Since typical eukaryotic cells have 10 times as many genes as do bacteria, the whole process of transcription and its regulation is more complex. For a start, eukaryotes have three different RNA polymerases, unlike bacteria which have just one. The three RNA polymerases transcribe different categories of nuclear genes. In addition, mitochondria and chloroplasts have their own RNA polymerases, which resemble the bacterial enzyme.

allosteric protein Protein that changes shape when it binds a small molecule

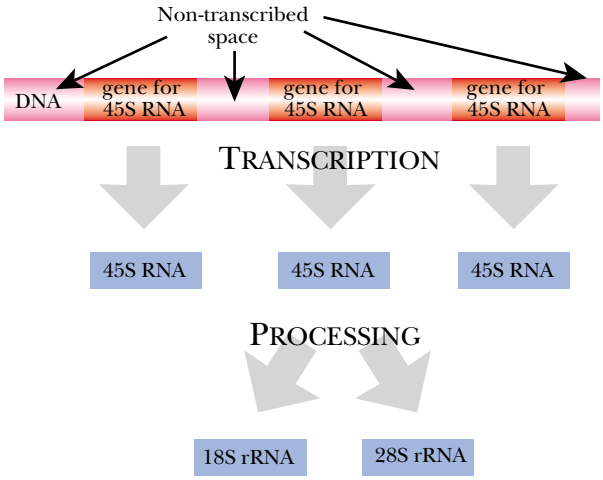


FIGURE 6.16 ***Clusters of Ribosomal RNA Genes***

The genes for rRNA are located at multiple sites along the DNA. A single transcribed unit of DNA yields an initial RNA molecule of 45S. The 45S RNA is processed to yield the final 18S and 28S subunits of rRNA.

RNA polymerase I transcribes the genes for the two large ribosomal RNA molecules and **RNA polymerase III** transcribes the genes for tRNA, 5S rRNA and a few other small RNA molecules. **RNA polymerase II** transcribes most eukaryotic genes that encode proteins and as a result is subject to the most complex regulation. Since ribosomal RNA and transfer RNA are needed all the time by all types of cells, RNA polymerases I and III operate constitutively in most cell types.

Many transcription factors are involved in controlling gene expression in eukaryotes.

A variety of proteins, known as **transcription factors** are also needed for the correct functioning of RNA polymerases. Transcription factors may be divided into general transcription factors and specific transcription factors. General transcription factors are needed for the transcription of all genes transcribed by a particular RNA polymerase, and are typically designated TFI, TFII, TFIII followed by individual letters. The I, II, and III refer to the corresponding RNA polymerase (see below). Specific transcription factors are needed for transcription of particular specific gene(s) under specific circumstances. [Proteins such as the sigma subunit of bacterial RNA polymerase may also be regarded as transcription factors, however, this terminology is usually only used for eukaryotes.]

Transcription of rRNA and tRNA in Eukaryotes

Eukaryotes contain many copies of the genes for ribosomal RNA. These are found in clusters and are transcribed by RNA polymerase I.

The genes for the two large ribosomal RNAs are present in multiple copies, from seven in *E. coli* to several hundred in higher eukaryotes (Fig. 6.16). In bacteria, the copies are dispersed, but in eukaryotes they form clusters of tandem repeats. In humans, there are clusters of rRNA genes on five separate chromosomes. The 18S and the 28S rRNA are transcribed together as a single large RNA (45S RNA) that is cleaved to release the two separate ribosomal RNA molecules. Between these transcription units are non-transcribed spacer regions. In eukaryotic cells, the rRNA genes have their own RNA polymerase to transcribe them, RNA polymerase I.

Synthesis of rRNA is localized to a special zone of the nucleus known as the nucleolus. Here the rRNA precursor is both transcribed and processed into 18S and 28S rRNA. These rRNA molecules then bind proteins, giving ribonucleo-protein particles. This yields a dense granular region when seen under the microscope (Fig. 6.17). The segments of chromosomes associated with the nucleolus were named "**nucleolar organizers**." It is now known that these correspond to the clusters of rRNA genes.

nucleolar organizer Chromosomal region associated with the nucleolus; actually a cluster of rRNA genes
RNA polymerase I Eukaryotic RNA polymerase that transcribes the genes for the large ribosomal RNAs
RNA polymerase II Eukaryotic RNA polymerase that transcribes the genes encoding proteins
RNA polymerase III Eukaryotic RNA polymerase that transcribes the genes for 5S ribosomal RNA and transfer RNA
transcription factor Protein that regulates gene expression by binding to DNA in the control region of the gene

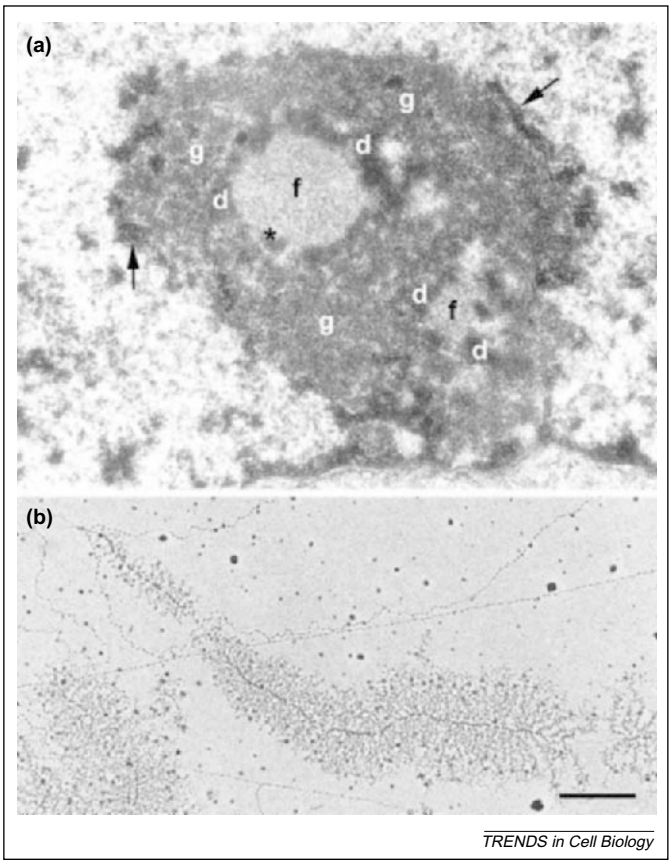


FIGURE 6.17 ***Ribosomal RNA is Made in the Nucleolus***

A) Electron micrograph of a thin-sectioned nucleolus from a mouse cell fixed in situ. Black arrows indicate peri-nucleolar condensed chromatin and the asterisk shows dense fibrillar components (d) clumping around fibrillar centers (f). Granular regions (g) of newly made ribonucleoproteins are also marked. Image provided by Ulrich Scheer, University of Würzburg. B) Spread Christmas tree structure (4 microns long) from a mouse cell is shown at the same magnification as (A). Bar represents 0.5 micron. From: Raska I., Oldies but goldies: searching for Christmas trees within the nucleolar architecture. Trends in Cell Biology 13 (2003) 517–525.

Although most promoters are AT rich, presumably because the weaker base pairs help in opening up the DNA, the promoter for RNA polymerase I is unusual in containing many GC pairs. There are two GC-rich regions, the core promoter and the upstream control element, that are 80 to 90 percent identical in sequence (Fig. 6.18). Both are recognized by protein UBF1 (Upstream Binding Factor 1), a single polypeptide. After UBF1 has bound, another protein, selectivity factor SL1, binds next to it. SL1 consists of four polypeptides, one of which, TBP (TATA Binding Protein), is also required for RNA polymerases II and III (see below). Once UBF1 and SL1 are in place, RNA polymerase I can bind. It is uncertain how the binding of UBF1 and SL1 at the upstream control element helps initiation in the case of RNA polymerase I. However, in similar cases, the DNA is known to bend around, bringing the upstream element into direct contact with the promoter region.

RNA polymerase III transcribes genes for small non-coding RNAs, in particular tRNA and 5S rRNA.

RNA polymerase III is responsible for making 5S rRNA and transfer RNA. It also makes some small nuclear RNAs, while other snRNAs are transcribed by RNA polymerase II (see below). The promoters for 5S rRNA and tRNA are unique and somewhat bizarre in being internal to the genes. Transcription of these genes requires the binding of either of two proteins known as TFIIIA and TFIIIC to a region over 50 bp downstream from the start site (Fig. 6.19). Once these have bound, they enable TFIIIB to bind to the region around the start of transcription. TFIIIB consists of three polypeptides, including TBP, and positions RNA polymerase III correctly at the start site.

As the promoters for RNA polymerase I and RNA polymerase III illustrate, recognition factor sites may be upstream or downstream from the start of transcription. However, in both cases, a positioning factor (SL1 or TFIIIB, respectively) is required to make sure that the polymerase starts transcribing at the correct place. These positioning factors thus play a similar role to that of the sigma factor in bacteria.

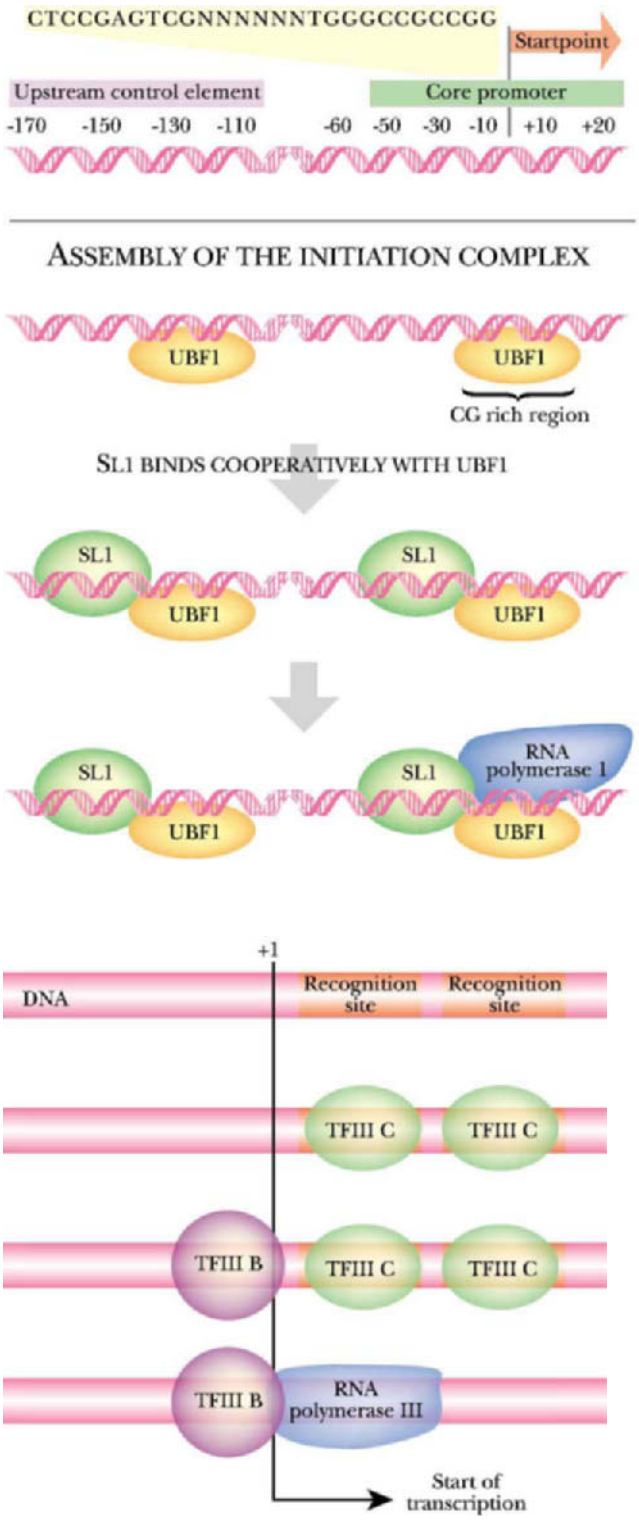


FIGURE 6.18 ***RNA Polymerase I Transcribes rRNA Genes***

The promoter for RNA polymerase I has an upstream control element and a core promoter, the latter rich in GC sequences. The UBF1 protein recognizes and binds to both the upstream control element and the core promoter. Subsequently, SL1 binds to the DNA in association with UBF1. Finally, RNA polymerase I binds and transcription commences. How this binding pattern facilitates transcription of rRNA is not known.

FIGURE 6.19 ***Internal Promoter for RNA Polymerase III***

The gene for 5S rRNA is transcribed using a promoter located within the gene itself. The recognition sites are downstream of the start site. TFIIIC (or TFIIIA) binds to both sites and this induces TFIIIB to bind to the promoter near the start site. Only after TFIIIB binds can RNA polymerase III bind.

Transcription of Protein-Encoding Genes in Eukaryotes

RNA polymerase II transcribes genes that code for proteins.

RNA polymerase II transcribes most eukaryotic genes that encode proteins. Recognition of the promoter and initiation of transcription by RNA polymerase II requires a number of general transcription factors. In addition, since many protein-encoding genes vary markedly in expression, a variety of specific transcription factors are needed for expression of certain genes under particular circumstances. For example, in a multi-

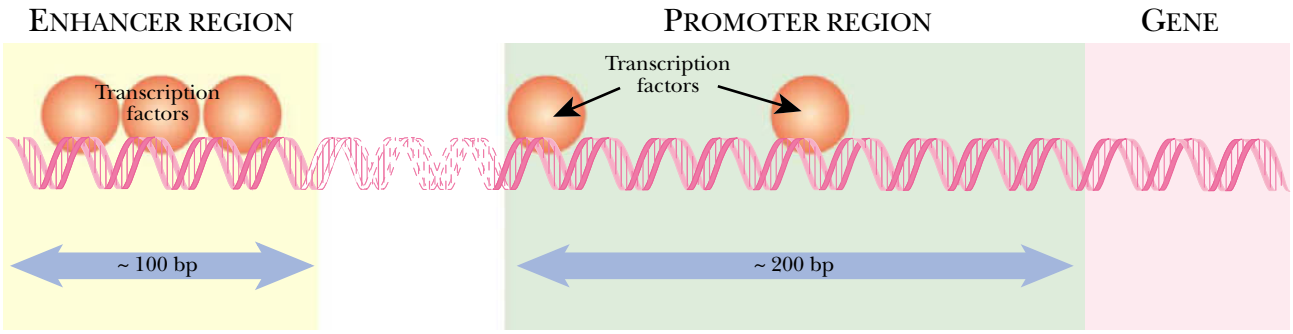


FIGURE 6.20 ***Promoter and Enhancer***

Although one RNA polymerase is used to transcribe most protein encoding genes, specificity is controlled by transcription factors and their recognition sequences. The promoter region is close to the start site and usually binds several transcription factors. In addition, extra transcription factors bind to regions known as enhancers. These may be far upstream of the promoter, as shown, or may be located downstream. Binding of the transcription factors to their recognition sequences influences polymerase activity and gene expression.

cellular organism, different cell types produce different types of proteins. Thus, red blood cells produce hemoglobin, whereas white blood cells make antibodies. Further, protein production often varies during development. Fetal hemoglobin is different from the adult version.

Some transcription factors bind to the promoter region, others to distant enhancer sequences.

The assorted transcription factors bind to and recognize specific sequences on the DNA. These DNA sequences are of two major classes, those comprising the promoter itself and a variety of **enhancer** sequences (Fig. 6.20). The general transcription factors for RNA polymerase II (TFII factors) bind to the promoter region. However, although some of the specific transcription factors also bind to the promoter region, others bind to the enhancer.

In eukaryotes, many protein-encoding genes are interrupted by introns. These are removed at the RNA stage. Consequently, transcription of DNA to give RNA does not yield messenger RNA directly. The RNA that results from transcription is known as the **primary transcript** and must be processed as described in Chapter 12 to give mRNA. The present discussion will therefore be limited to the transcription of genes by RNA polymerase II to give the primary transcript.

The TATA box is the critical sequence that allows RNA polymerase II to recognize the promoter.

Promoters for RNA polymerase II consist of three regions, the **initiator box**, the **TATA box** and a variety of **upstream elements** (see below). The initiator box is a sequence found at the site where transcription starts. The first transcribed base of the mRNA is usually A with a pyrimidine on each side, as in bacteria. The consensus is weak: YYCAYYYYY (where Y is any pyrimidine). About 25 base pairs upstream from this is the TATA box, an AT-rich sequence, which is recognized by the same factor TBP (**TATA binding protein** or **TATA box factor**) that is needed for binding of RNA polymerases I and III. TBP is unusual in binding in the minor groove of DNA. (Almost all DNA-binding proteins bind in the major groove). On both sides of the TATA box are GC-rich regions (Fig. 6.21).

TBP is found in three different protein complexes, depending on whether RNA polymerase I, II or III is involved. In the present case, TBP forms part of a transcription factor complex known as TFIID that is needed to recognize promoters specific for RNA polymerase II. Several other TFII complexes are also needed for RNA polymerase II function. TFIIA and TFIIB bind next. Then at last RNA polymerase II itself

enhancer Regulatory sequence outside, and often far away from, the promoter region that binds transcription factors
initiator box Sequence at the start of transcription of a eukaryotic gene
primary transcript RNA molecule produced by transcription before it has been processed in any way
TATA binding protein (TBP) Transcription factor that recognizes the TATA box
TATA box Binding site for a transcription factor that guides RNA polymerase II to the promoter in eukaryotes
TATA box factor Another name for TATA binding protein
upstream element DNA sequence upstream of the TATA box in eukaryotic promoters that is recognized by specific proteins

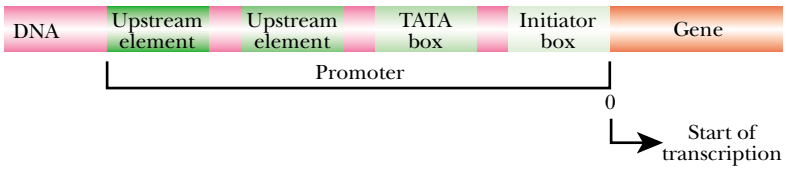


FIGURE 6.21 ***Eukaryotic Promoter Components—Initiator and TATA Boxes***

The promoter for RNA polymerase II has an initiator box at the start site and a TATA box slightly upstream of this. Further upstream there are normally several upstream elements (two are shown here).

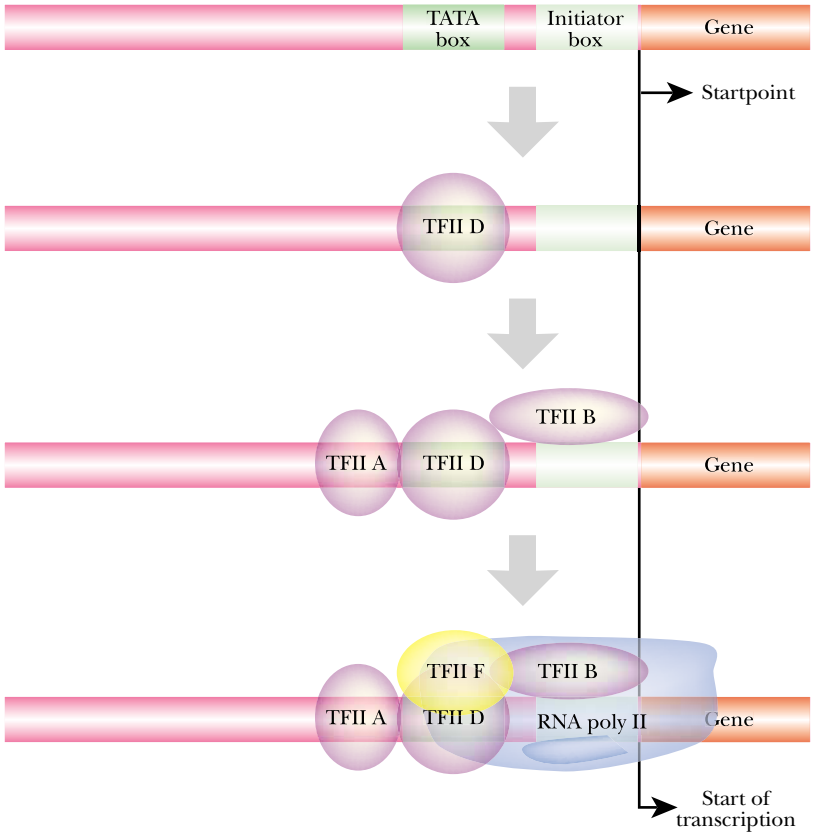


FIGURE 6.22 ***Binding of RNA Polymerase II to Promoter***

Starting with TFIID, which contains TATA binding protein, the components of the TFII complex bind one after another. Finally TFIIF helps RNA polymerase II to bind to the DNA.

arrives, accompanied by TFIIF which probably helps RNA polymerase bind (Fig. 6.22). At this point RNA polymerase II can initiate synthesis of RNA. However, it is not yet free to move away from the promoter.

Release of RNA polymerase II from the promoter and elongation of the RNA requires three more TFII complexes, TFIIE, TFIIH and TFIIJ. In particular, TFIIH must phosphorylate the tail of RNA polymerase before it can move (Fig. 6.23). The tail, or **CTD (carboxy-terminal domain)**, consists of a seven-amino acid sequence (Tyr Ser Pro Thr Ser Pro Ser) repeated approximately 50 times. This may be phosphorylated on the serine or threonine residues. All of the TFII complexes except for TFIIH are left behind as RNA polymerase moves forward.

Like bacterial RNA polymerase, the eukaryotic RNA polymerases all have multiple subunits. RNA polymerase II has more than 10 subunits and shares three of these

CTD (carboxy-terminal domain) Repetitive region at the C-terminus of RNA polymerase II that may be phosphorylated

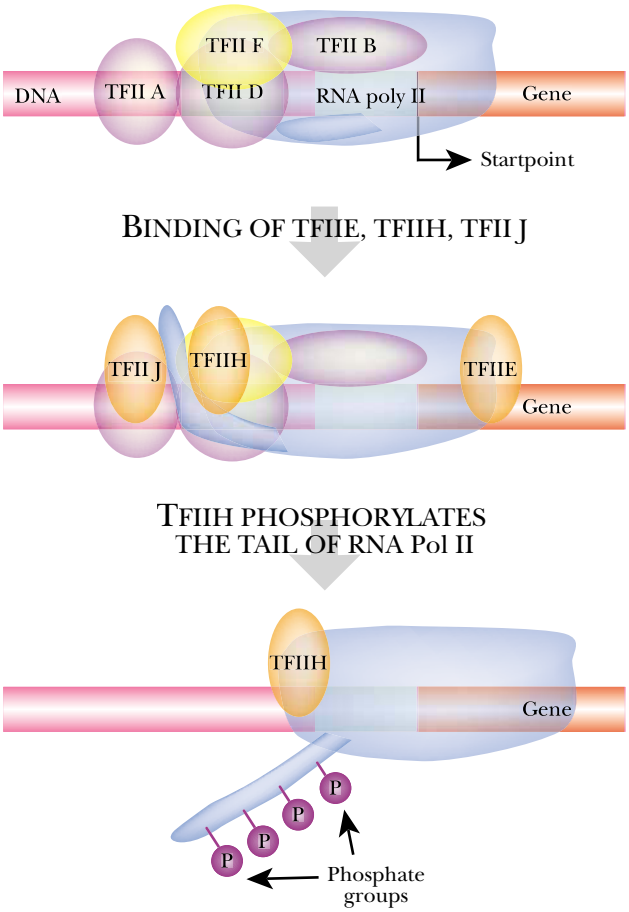


FIGURE 6.23 ***RNA Polymerase II Moves Forward from the Promoter***

Before RNA polymerase II can move forward, the binding of other factors must occur. One of these, TFIIH, phosphorylates the tail of RNA polymerase II. The tail changes position with respect to the body of RNA polymerase II. The other factors leave and RNA polymerase moves along the DNA and begins the process of transcription.

with RNA polymerases I and III. The largest subunit of RNA polymerase II is related to the β′ subunit of bacterial RNA polymerase and possesses the CTD tail. In addition, the assorted TFII complexes each consist of several polypeptide chains. Thus the initiation complex for RNA polymerase II includes over 20 polypeptides.

Upstream Elements Increase the Efficiency of RNA Polymerase II Binding

Upstream elements close to the promoter bind a range of specific transcription factors.

RNA polymerase II can bind and initiate transcription at minimal promoter consisting of just an initiator and TATA boxes. However, this is extremely inefficient unless upstream elements are also present. There are many different upstream elements. They are typically five to ten base pairs long and located from 50 to 200 bases upstream of the start site. There may be more than one upstream element in a given promoter and the same upstream element may be found at different places in different promoters.

The TFII proteins are *general* transcription factors because they are always required. In contrast, *specific* transcription factors affect only certain genes and are involved in regulating gene expression in response to a variety of signals (Fig. 6.24). The upstream elements are the recognition sites for specific transcription factors. These usually make contact with the transcription apparatus via TFIID, TFIIB or TFIIA, not by directly touching RNA polymerase II itself. Most commonly, binding is to TFIID. Binding of the specific transcription factors helps assembly of the transcription apparatus and therefore increases the frequency of initiation.

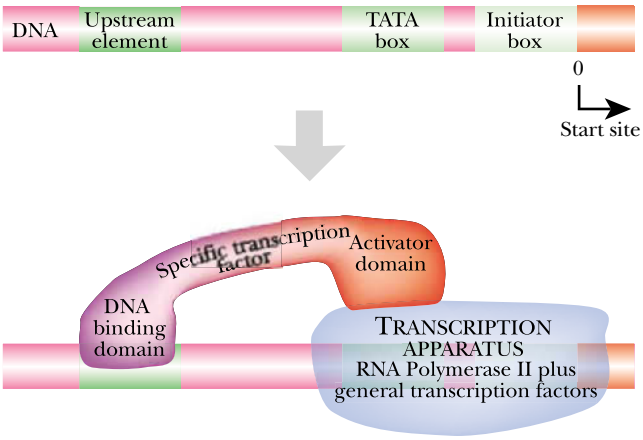


FIGURE 6.24 ***Upstream Elements Facilitate Transcription***

The upstream elements make contact with one domain of an activator protein. The activator also binds to the transcription apparatus near the start site.

TABLE 6.01 General Transcription Factors Associated with RNA Polymerase II

TBP	binds to TATA box, part of TFIID
TFIID	includes TBP, recognizes Pol II specific promoter
TFIIA	binds upstream of TATA box; required for binding of RNA Pol II to promoter
TFIIB	binds downstream of TATA box; required for binding of RNA Pol II to promoter
TFIIF	accompanies RNA Pol II as it binds to promoter
TFIIE	required for promoter clearance and elongation
TFIIH	phosphorylates the tail of RNA Pol II, retained by polymerase during elongation
TFIIJ	required for promoter clearance and elongation

In eukaryotes most genes are subject to positive control. Repressors are rare and usually behave differently to those in bacteria.

Common upstream elements include the GC box, CAAT box, AP1 element and Octamer element. The GC box (GGGCGG) is often present in multiple copies. Despite being nonsymmetrical, the GC box works in either orientation and is recognized by the SP1 factor. Some upstream elements are recognized by more than one protein. In these cases, different transcription factors are often present in different tissues. For example, the Oct-1 and Oct-2 proteins both recognize the Octamer element. Oct-1 is found in all tissues but Oct-2 only appears in immune cells, where it helps activate genes encoding antibodies.

The specific transcription factors that bind to the upstream elements and enhancers are thus gene activator proteins. Repressors are rare in eukaryotes. Furthermore, when found, they do not bind to the DNA directly and block the binding of the RNA polymerase. Instead, they bind to some component of the growing transcription apparatus and block further assembly.

Enhancers Control Transcription at a Distance

Enhancer sequences are located far away from the genes they control.

Enhancers are sequences that are involved in gene regulation, especially during development or in different cell types. Enhancers do exactly what their name indicates—they enhance the initiation of transcription as a result of binding specific transcription factors. Enhancers often consist of a cluster of recognition sites and therefore bind several proteins. Some recognition sites (e.g., Octamer and AP1) are found in both enhancers and as upstream elements in promoters.

Although enhancers are sometimes close to the genes they control, more often they are found a considerable distance, perhaps thousands of base pairs away, not even

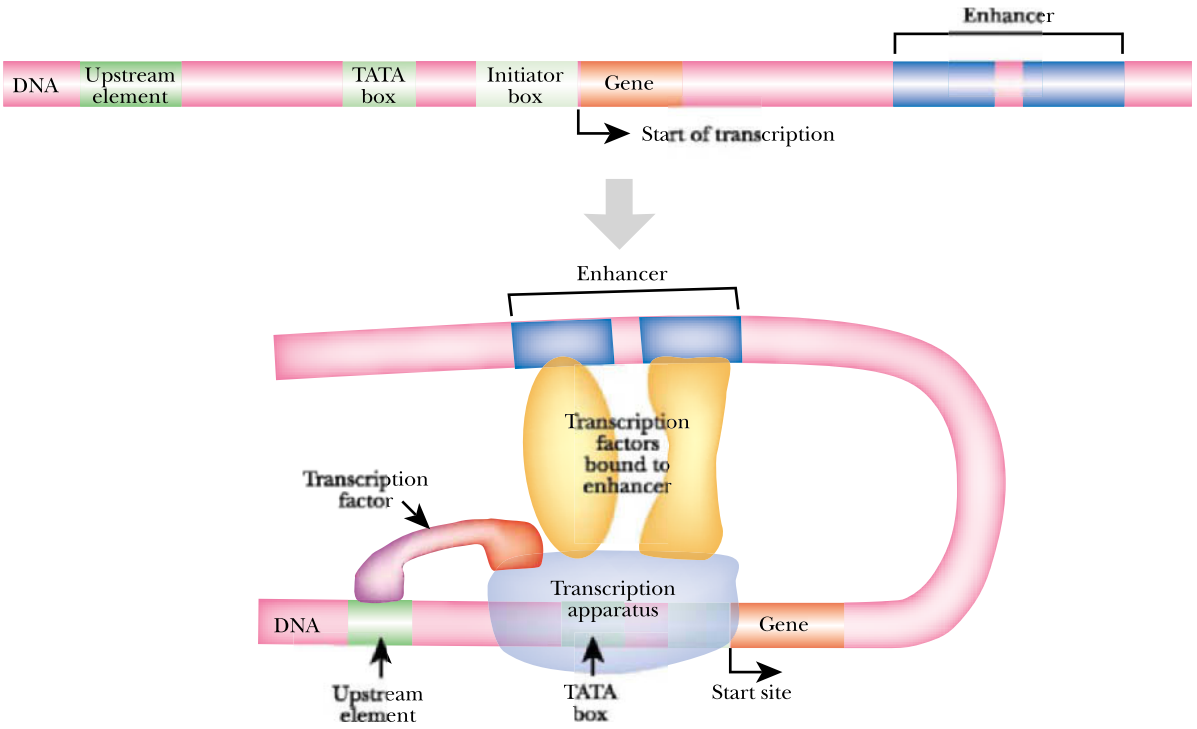


FIGURE 6.25 ***Looping Model for Enhancer***

The enhancer shown here is located downstream of the start site. To enhance transcription, the enhancer first binds several transcription factors. Subsequently, the DNA forms a loop allowing the enhancer to make contact with the transcription apparatus via the bound transcription factors.

associated with the gene. Enhancers may be located either upstream or downstream from the promoter and the position may vary from case to case. In addition, enhancers function equally well in either orientation. Experiments in which enhancers have been moved have shown that an enhancer will increase transcription from any promoter within its neighborhood. These properties imply that the enhancer must make contact with the transcription apparatus. When an enhancer switches a gene on, the DNA between it and the promoter loops out as shown in Figure 6.25.