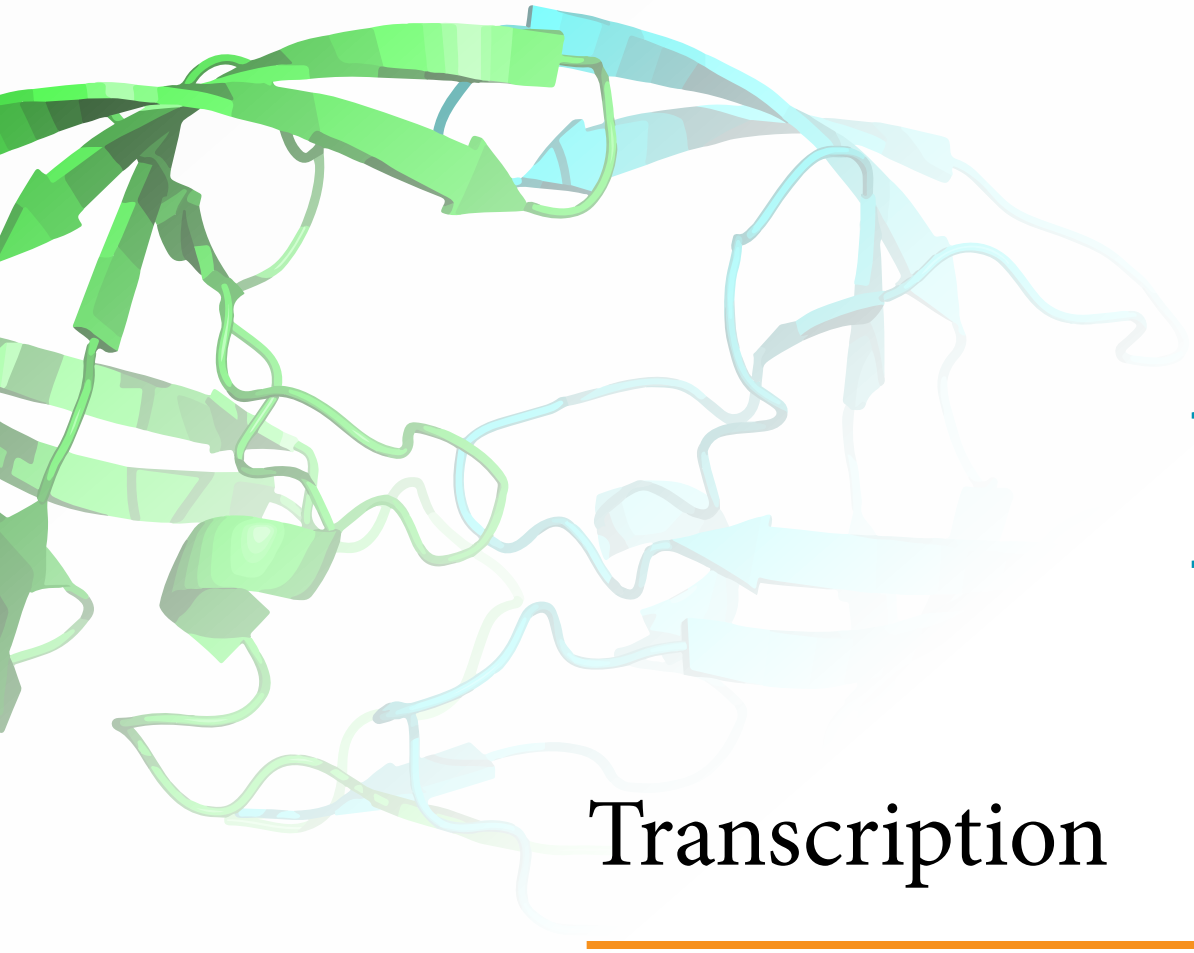

10

Transcription

Goal

To understand how genetic information is transmitted from the genome to the ribosome.

Objectives

After this chapter, you should be able to

- compare and contrast RNA with DNA.
- compare and contrast transcription with replication.
- explain the differences between the transcription machinery of bacteria and eukaryotes.
- describe how transcripts are processed into mRNAs.
- diagram the transesterification reactions that mediate splicing.

The exquisite structure of the double helix provided a simple explanation for how DNA serves as an information carrier and for how it can be replicated. Now we turn our attention to the question of how genetic information in the order of base pairs in DNA is transmitted from the double helix to the ribosome, where it is translated into sequences of amino acids. Thus, here we focus on the structure and properties of **ribonucleic acid (RNA)**, how it is copied from DNA, and how it is processed into a form known as messenger RNA that can direct the synthesis of proteins.

RNA contains a 2' OH on its sugars and uracil in place of thymine

Like DNA, RNA is an alternating copolymer of phosphates and sugars. Unlike DNA, however, the RNA backbone is composed of ribose sugars rather than 2'-deoxyriboses. Ribose contains a hydroxyl group at the 2' position in place of one of the two hydrogens present in 2'-deoxyribose (Figure 1A). Also, and importantly, thymine in DNA is replaced by uracil in RNA. Like thymine, uracil is a pyrimidine and, like thymine, it base pairs with adenine. Uracil differs from thymine only in the absence of a methyl group on carbon 5 (Figure 1B). Thymine is thus more energetically costly for the cell to produce than uracil. Indeed, thymine arises from the methylation of uracil. Why then do cells bother to have thymine in DNA? As we speculated in the previous chapter, cells likely invest in the extra methylation step as a strategy for distinguishing thymine in the genetic material from uracil arising from the deamination of cytosine.

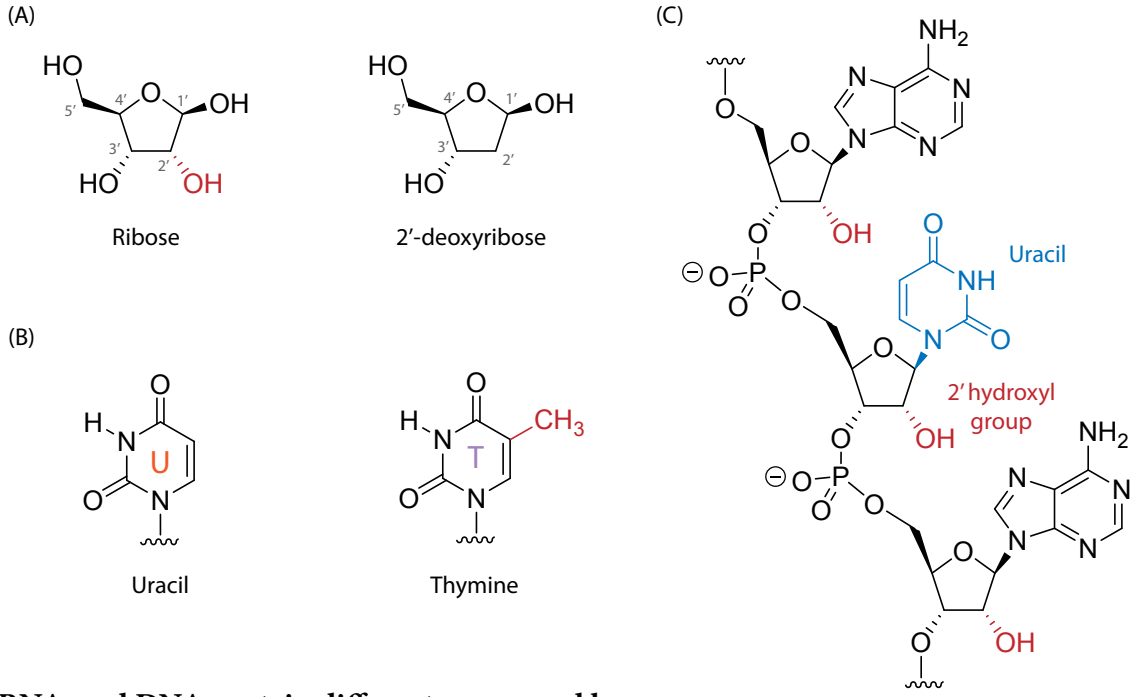

Figure 1 RNA and DNA contain different sugars and bases

Shown in (A) is a comparison of the structures of ribose and 2'-deoxyribose and in (B) of thymine and uracil. Shown in (C) is a short stretch of RNA.

Other than these differences, RNA and DNA are chemically identical. Nonetheless, as we will see in subsequent chapters, RNA is more versatile than DNA. DNA functions exclusively as an information carrier, and it is generally restricted to a single structure, the double helix. RNA, in contrast, can form complex three-dimensional structures of many kinds and can perform multiple functions in the cell, including acting as a catalyst. Here, however, we are principally concerned with its role in transmitting sequence information from DNA to the ribosome.

The 2' hydroxyl contributes to RNA's versatility, particularly as a catalyst, as we will see in Chapter 13. At the same time, the 2' hydroxyl imparts a cost to RNA, rendering it less stable than DNA and prone to self-cleavage, as explained in Box 1. Indeed, this effect on stability may explain why DNA lacks a 2' hydroxyl, as stability would be expected to be at a premium for DNA's role as an information repository.

Box 1 The 2' hydroxyl group renders RNA susceptible to auto cleavage

RNA is more challenging to work with in the laboratory than is DNA because it readily breaks down into smaller fragments, especially at elevated pH. The basis for this instability is the 2' hydroxyl, which promotes an auto cleavage reaction. In this reaction, the oxygen of the 2' hydroxyl with its non-bonded lone pairs acts as a nucleophile, attacking the phosphorus atom of the adjacent 3' phosphate group. As a consequence, a phosphodiester bond is formed between 2' and 3' hydroxyls, resulting in a cyclic product and scission of the 3'-to-5' phosphodiester bond that linked the ribose to the adjacent ribose in the polynucleotide. The reaction is more favorable at high pH because elevated levels of OH^- facilitate deprotonation of the 2' OH group. Also, each and every phosphodiester bond in the polynucleotide chain of RNA is susceptible to this auto cleavage reaction.

This auto cleavage reaction is instructive from a mechanistic point of view. Even though the negative charge surrounding the phosphate group helps to protect the phosphorus atom from attacking water molecules

(as we saw for DNA hydrolysis; Chapter 8, Box 3), the 2' oxygen atom of ribose is a particularly potent nucleophile. Unlike the oxygen atom of a freely diffusing water molecule, the 2' oxygen atom is held close to the phosphorus atom and hence, in effect, there is a high local concentration of the nucleophile that is in a favorable alignment with the phosphorus. (This high-local-concentration effect is analogous to the effects of cooperativity on DNA annealing considered in Chapter 8.)

We refer to the RNA auto cleavage reaction as an **intramolecular** reaction because it involves two functional groups within the same molecule as opposed to an **intermolecular** reaction, which involves a reaction between functional groups on different molecules (Figure 2). Intramolecular reactions tend to take place much faster than intermolecular reactions because the reacting groups are tethered to each other, as in the case of RNA.

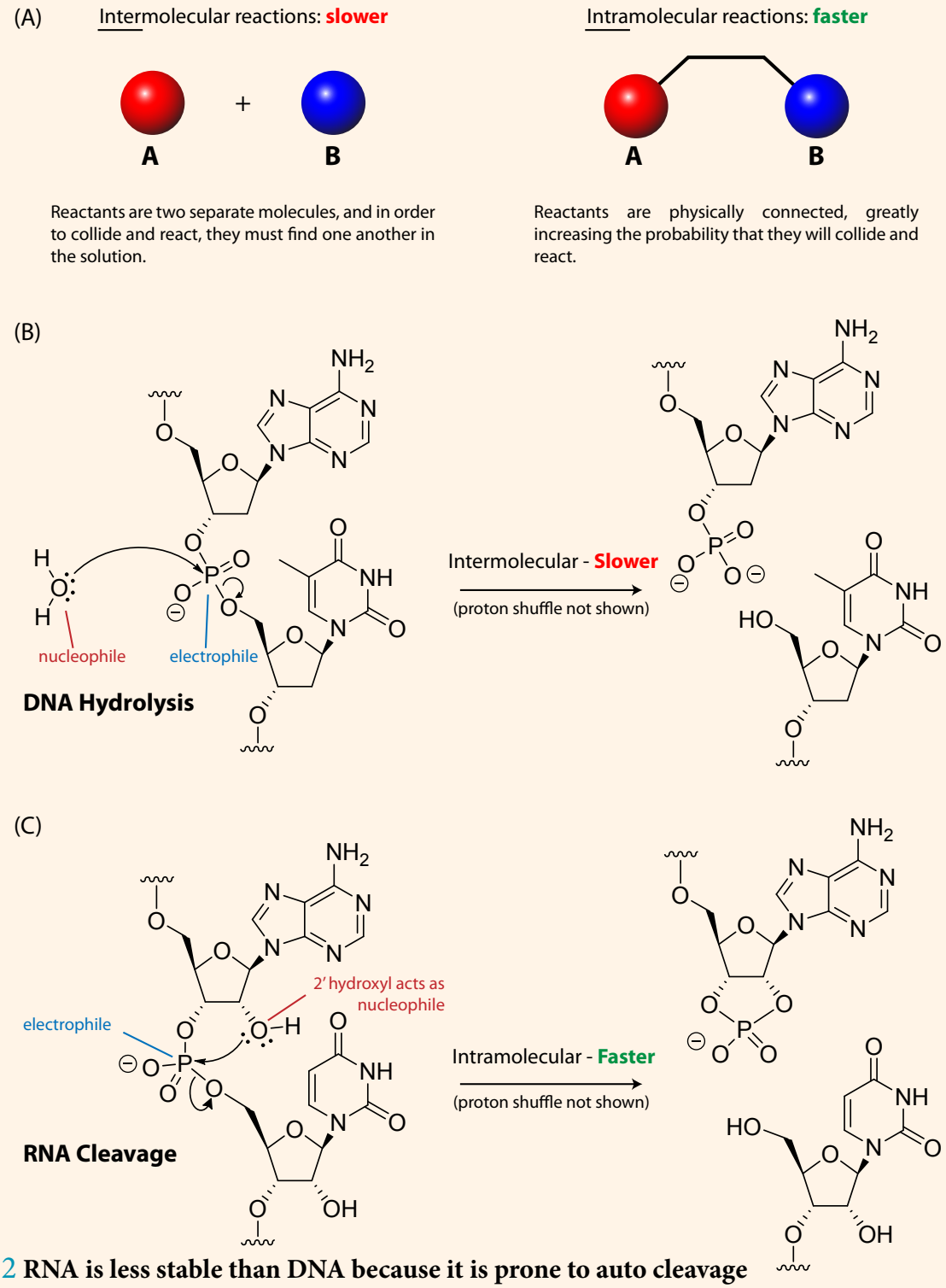

Figure 2 **RNA is less stable than DNA because it is prone to auto cleavage**

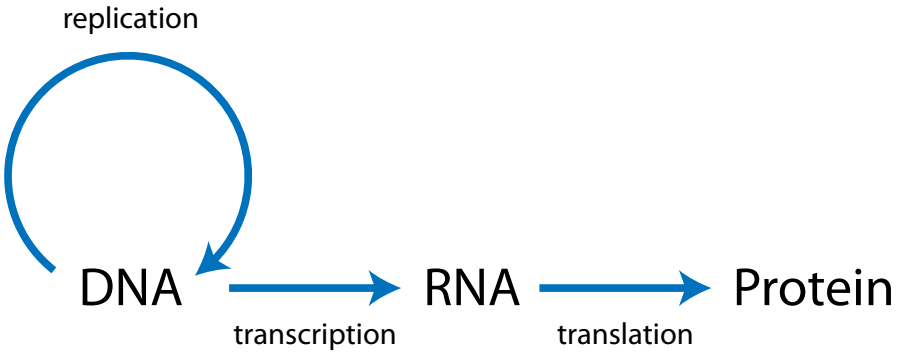

Figure 3 **The central dogma describes how genetic information is transferred among DNA, RNA, and proteins**

The central dogma states that information flows from nucleic acid to protein

The overarching tenet of molecular biology is that information in the form of the order of bases and in the form of the order of amino acids flows from nucleic acid to nucleic acid and from nucleic acid to protein but not back again. This tenet was enunciated by Francis Crick in 1958 as the central dogma:

> *"The central dogma states that once 'information' has passed into protein it cannot get out again. The transfer of information from nucleic acid to nucleic acid, or from nucleic acid to protein, may be possible, but transfer from protein to protein, or from protein to nucleic acid, is impossible. Information means here the precise determination of sequence, either of bases in the nucleic acid or of amino acid residues in the protein."*

Somehow information in the form of the linear sequence of bases must be conveyed to the protein synthesis machinery of the cell, the ribosome, which is the subject of the next chapter. This requires an intermediary, as the double helix cannot and does not directly interact with ribosomes. Indeed, in eukaryotic organisms, DNA is sequestered in the nucleus of the cell, whereas protein synthesis takes place in the cytoplasm. As we have already indicated, the intermediary is RNA or, more specifically, messenger RNA (mRNA), which is copied from one of the two strands of the double helix corresponding to a gene or small group of genes before being transmitted to the ribosome. The process by which a stretch of DNA is copied into messenger RNA is called **transcription**, and the process by which messenger RNA is used to direct the synthesis of protein on the ribosome is called **translation**.

Figure 3 encapsulates the central dogma in showing that information in the form of DNA (nucleic acid) can be transmitted to DNA (nucleic acid) in DNA replication and to protein via the intermediary of RNA and the processes of transcription and translation.

DNA is transcribed asymmetrically in a moving bubble

During transcription the two strands of the double helix are temporarily unwound (by an enzyme known as RNA polymerase as we will come to) to create a **transcription bubble** that is approximately 13 base pairs in length

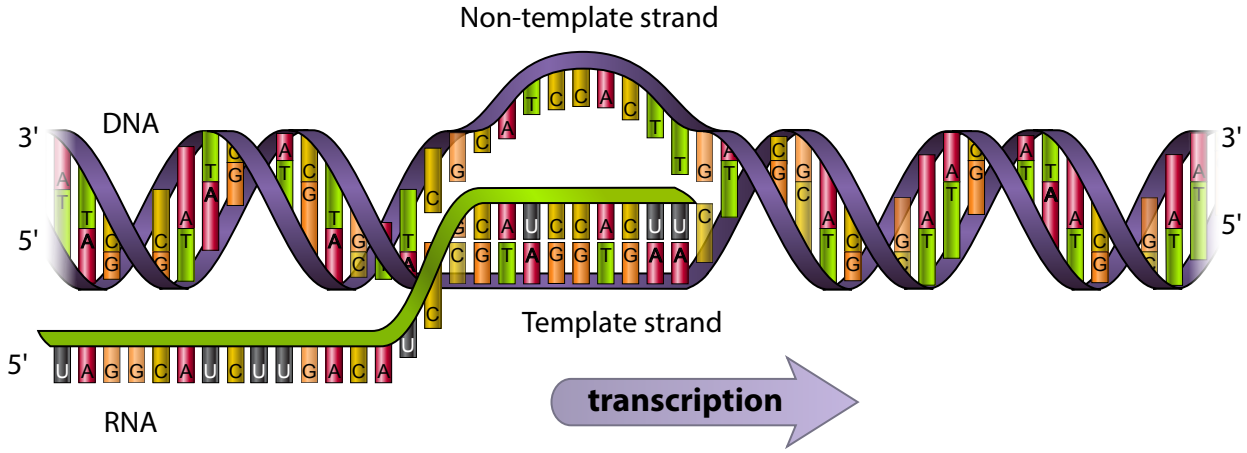

Figure 4 **RNA is synthesized at a moving transcription bubble**

(Figure 4). The two strands of the bubble are known as the **template** and the **non-template** strand. RNA is copied from the template strand. The region of strand separation moves down the double helix with continual unwinding and rewinding of the two strands to create a moving bubble.

Note that the non-template strand has the same sequence as the RNA transcript. Note too that the direction of transcription—left to right as shown—is determined by the strand that is being copied as dictated by the 5'-to-3' rule and the anti-parallel rule. That is, if the lower strand, which has its 3' end on the left, is being copied, then the RNA must be being synthesized from left to right. The product of transcription, the RNA, is extruded from the template. Thus, the growing transcript is extruded from the moving bubble with the template and non-template strands re-annealing as the region of strand separation progresses along the double helix.

Specific regions of DNA are transcribed into RNA

Whereas the entire genome is copied in DNA replication, only specific portions of the genome are copied into RNA. Generally speaking, these regions contain the coding information for specific proteins and hence correspond to genes. That is, the process of transcription copies the coding sequence for one or more adjacent genes into RNA. DNA that is copied into RNA is known as a **transcription unit**. A transcription unit originates from a specific site on DNA, the initiation site, and ends at a termination site. Whereas the genome is replicated only once during the cell cycle, transcription units are transcribed into RNA multiple times, resulting in multiple copies of the same transcript.

Finally, note that not all transcription units have the same orientation (Figure 5). Some transcription units are transcribed from left to right as shown in the cartoon and others from right to left. Because the 5'-to-3' and anti-parallel rules demand that the direction of transcription be set by the strand that is being copied, some transcription units point to the right and others to the left. This means therefore that the identity of the template strand varies according to the orientation of each transcription unit. In other words, the template strand is the lower strand in Figure 5 for some units and the upper strand for others.

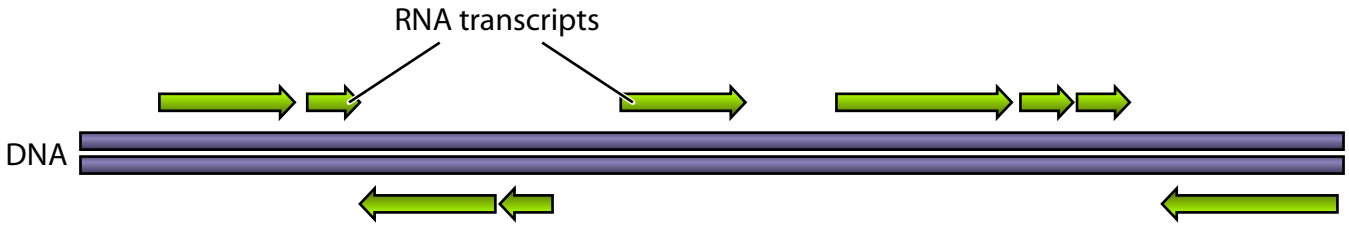

Figure 5 **Not all transcription units are transcribed in the same direction**

Transcription is catalyzed by RNA polymerase

The enzyme that catalyzes RNA synthesis is called **RNA polymerase**. As in DNA replication, the nucleotide sequence of the RNA chain is determined by base pairing between incoming nucleotides and the DNA template. When a match is made, the incoming ribonucleotide is covalently linked to the growing RNA chain in an enzymatically catalyzed reaction that proceeds at a rate of about 50 nucleotides per second (and is therefore more than an order of magnitude slower than the rate of DNA synthesis; Chapter 9). Like DNA polymerase, RNA polymerase uses **nucleoside triphosphates** (NTPs) as substrates. Specifically, RNA polymerase uses NTPs whereas, and as we have seen, DNA polymerase uses dNTPs. An important further difference is that whereas DNA polymerase uses dTTP, RNA polymerase uses UTP. As we have seen, uracil pairs with adenine. Therefore, adenine on the template strand is recognized by UTP in the same way it is recognized by dTTP in DNA synthesis. Unlike DNA polymerase, RNA polymerase does not have an editing pocket; since RNA is not the genetic material, transcription doesn't need to be as accurate as replication.

The transcription machinery differs significantly between bacteria and the cells of higher organisms. In what follows we will consider the bacterial RNA polymerase first and then that of higher cells.

Bacterial RNA polymerase consists of a core enzyme and a sigma subunit that recognizes start sites for transcription

RNA polymerase in bacteria is a heteromeric complex consisting of multiple protein subunits. The core enzyme, which is responsible for RNA synthesis, resembles a crab claw as can be seen in the X-ray crystallographic structure of Figure 6. The active site is at the base of the claw. At the start of transcription RNA polymerase binds to a DNA sequence called the **promoter** as we will explain. During binding to the promoter the DNA unwinds to create the transcription bubble (Figure 7). Next, transcription commences at the start site with double-stranded DNA downstream (that is, ahead) of the transcribing polymerase unwinding and template and non-template strands exiting the polymerase re-annealing into a double helix. Thus, as we have seen, RNA synthesis takes place in a moving transcription bubble. Notice that newly synthesized RNA is transiently in a DNA:RNA hybrid helix with the template strand but is then extruded from the bubble as upstream DNA re-anneals. During transcription the claw closes around the double helix, thereby helping to promote processivity.

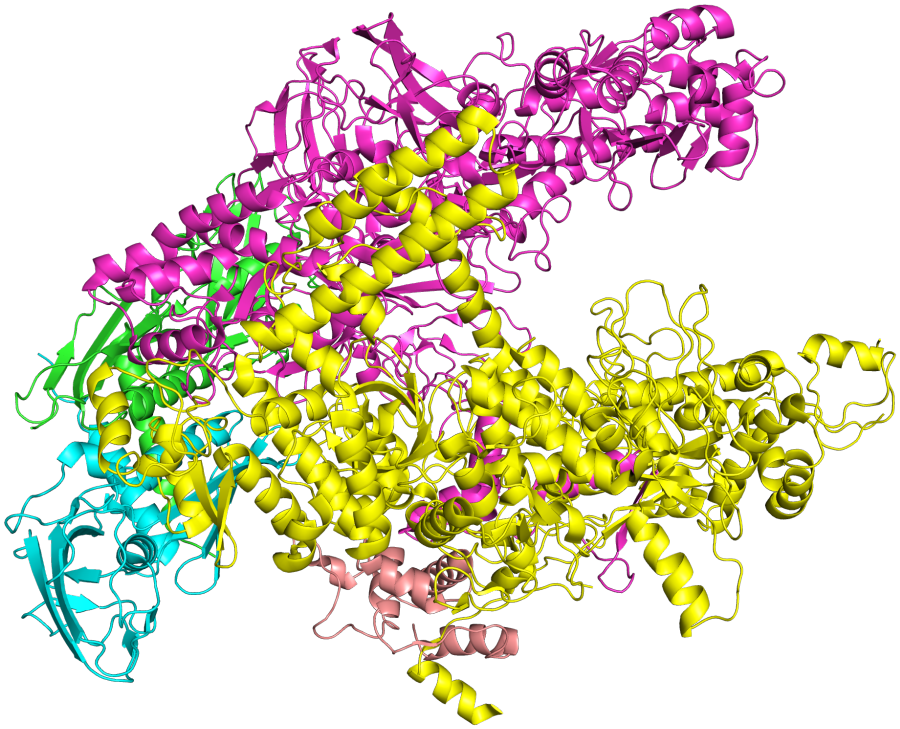

Figure 6 RNA polymerase resembles a "crab claw" with its active site at the base of the claw

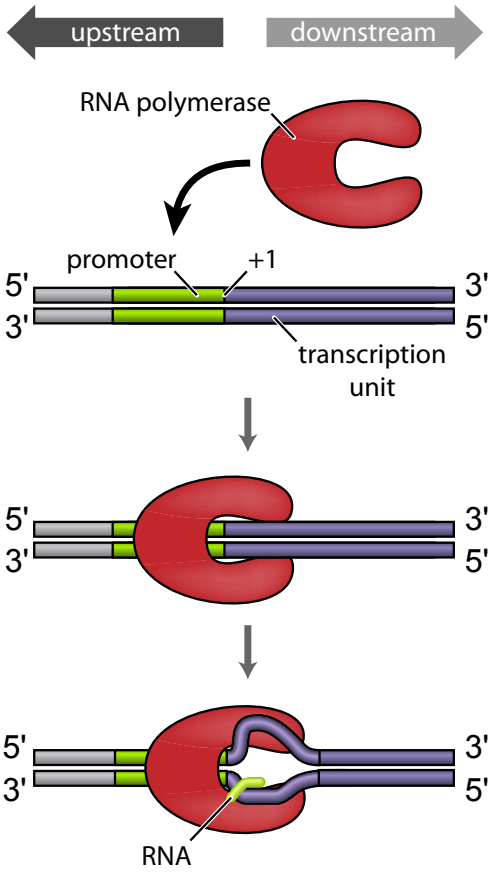

Figure 7 RNA polymerase binds to the promoter to initiate transcription

The RNA polymerase core enzyme is associated with an additional protein subunit called the **sigma factor**. The sigma factor is not involved in the polymerization of nucleotides. Instead, it directs the enzyme to promoter sites on the DNA where transcription will commence. Once transcription has begun, the sigma factor is released from the promoter and is no longer needed for continued copying of the transcription unit.

How does the sigma factor enable the RNA polymerase to commence transcription at specific sites on the DNA? The answer is that the sigma factor is a sequence-specific recognition protein that enables the enzyme complex to recognize and bind to the promoter, which lies just upstream of the start site. The promoter consists of two short stretches of DNA located roughly (although not exactly) 35 and 10 base pairs upstream of the start site that are referred to as the **–35** and **–10** sequences (Figure 8). The Platonic ideal for a –35 sequence is 5'-TTGACA-3' on the non-template strand. Likewise, the Platonic ideal for the –10 sequence is 5'-TATAAT-3', also on the non-template strand. These two sequences are separated from each other by a space of about 17 base pairs. Few promoters conform exactly to the Platonic ideal. Rather, they are approximations to it. The ideal –35 and –10 sequences represent a **consensus** obtained from comparing many different promoter sequences. In general, the closer the approximation to the ideal, the stronger the promoter is in the sense that RNA polymerase binds to it and initiates transcription more frequently.

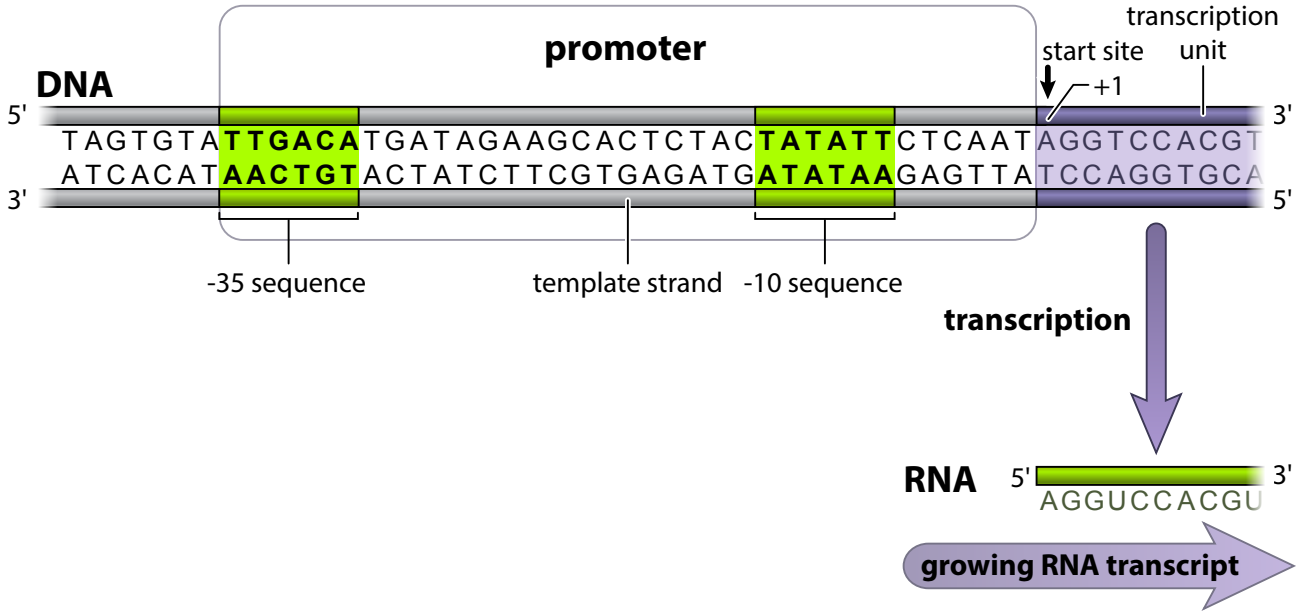

Figure 8 **Bacterial promoters contain –35 and –10 sequences that are recognized by sigma factors**

Box 2 Consensus sequences for promoters

Promoter sequences typically deviate from the ideal sequence that is recognized by the sigma factor. Figure 9 shows the relative frequency with which each nucleotide occurs at each location within the –35 and –10 sequences of many different promoters. As you can see, there is a consensus sequence of 5'-TTGACA-3' for the –35 sequence and 5'-TATAAT-3' for the –10 sequence, as those nucleotides occur most frequently at their respective locations. Note, however, that none of these nucleotides is absolutely conserved, and in fact many promoters deviate from this ideal sequence. Often times these deviations affect the relative ability of promoters to direct transcription, ultimately influencing the relative abundance of proteins in the cell. We will return to this idea in Chapter 12 when we will see an example of how a deviation from the ideal promoter sequence is used by the cell to regulate transcription. The concept of a consensus sequence applies generally to DNA-binding proteins. The specific sequence found at an individual binding site often differs from the consensus based on a comparison of many binding sites.

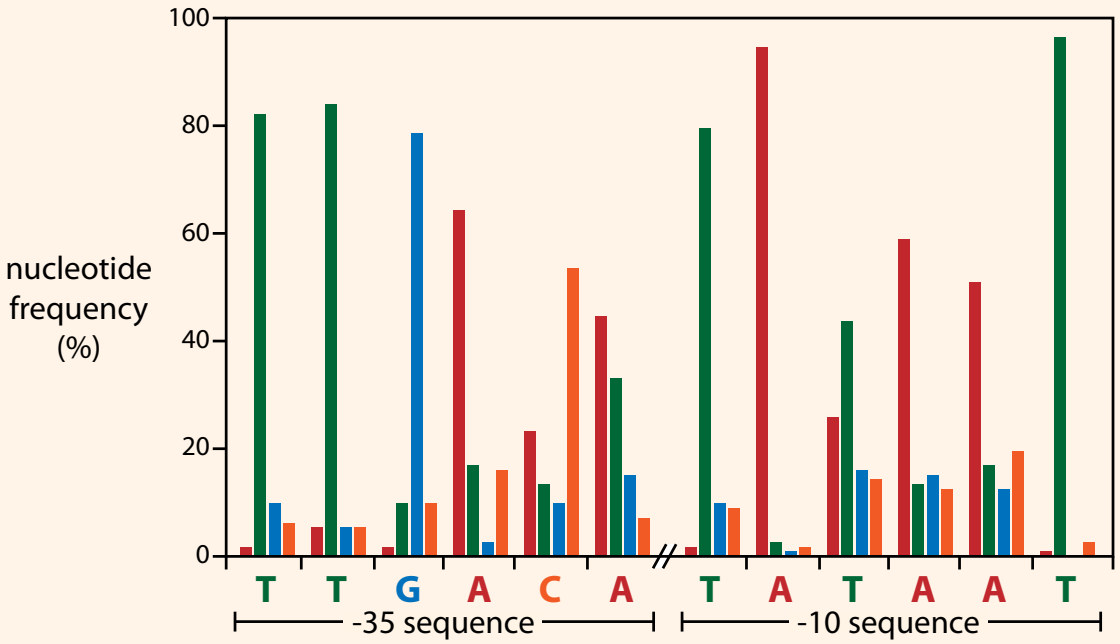

Figure 9 **Promoter sequences do not always match the ideal sequence**

Adapted from data reported in *Nucleic Acids Res.* 11:8 2237 (1983).

The sigma subunit directly recognizes and contacts the −35 and −10 sequence elements to facilitate RNA polymerase binding and the initiation of transcription. In the case of the −35 sequence, side chains of amino acids in the sigma subunit contact the edges of base pairs in the major groove, as we considered in Chapter 8. Interestingly, and in contrast to most DNA-binding proteins, key contacts between the sigma factor and the −10 element are made as the two strands separate during the binding of RNA polymerase to the DNA to create the transcription bubble (Figure 10).

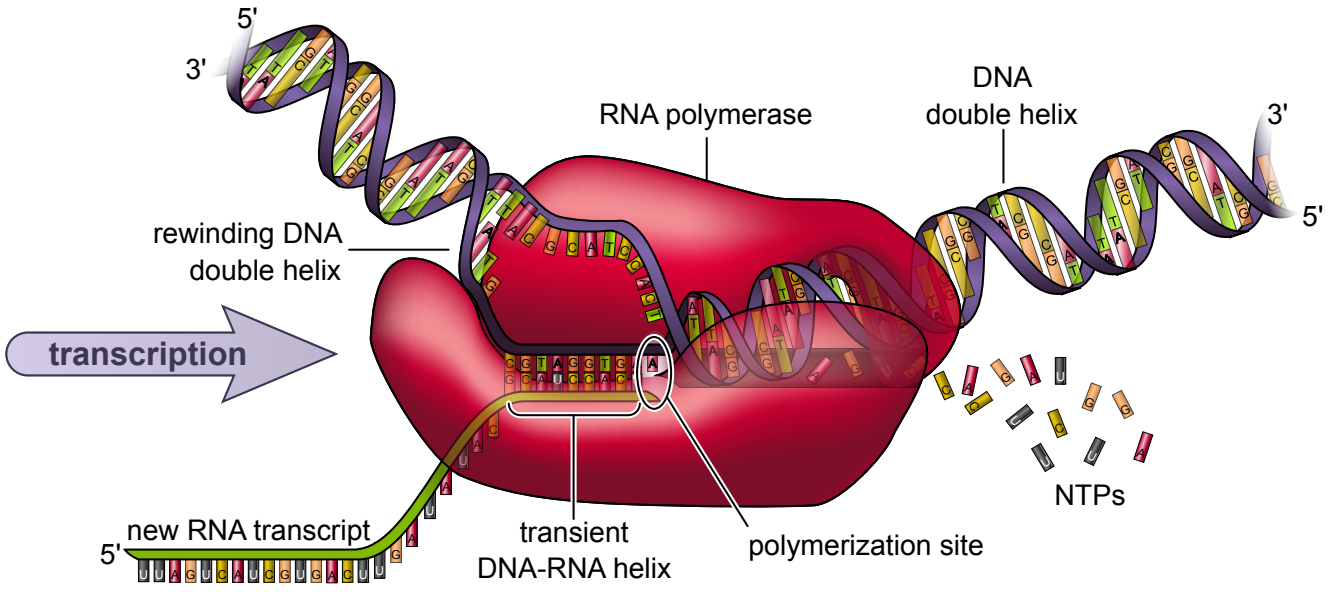

Figure 10 **Transcription is catalyzed by RNA polymerase**

Shown is RNA polymerase in the process of transcribing, that is, after the sigma factor (not shown) has directed the enzyme to the promoter.

The transcription machinery in cells of higher organisms is more complex and functions differently from that in bacteria

The process of transcription in the cells of higher organisms differs in many respects from that of bacteria. It is more complex, involving many more protein components and indeed multiple different RNA polymerases; the mode of promoter recognition, as we shall see, is fundamentally different; and the transcripts generated in eukaryotic cells undergo chemical modifications that are for the most part not observed in bacteria. Most conspicuously, eukaryotic cells have three different RNA polymerases, each responsible for transcribing different sets of genes. RNA polymerases I and III transcribe genes for various RNAs that do not encode amino acid sequences (so-called non-coding RNAs), such as the RNAs involved in the translation of messenger RNAs (e.g., tRNAs and ribosomal RNAs, which we will consider in the next chapter). Here we focus on RNA polymerase II, the enzyme that is responsible for generating messenger RNAs by transcribing protein-coding genes.

Promoter recognition in eukaryotic cells is mediated by proteins that bind to the DNA and recruit RNA polymerase

As we will see in Chapter 12, some promoters in bacteria and in higher organisms additionally depend on DNA-binding proteins that recognize other sites in DNA and assist in promoter recognition. Here we are simply concerned with the basic machinery for the recognition of promoters in eukaryotic cells. In bacteria, as we have seen, promoters are recognized by a protein, sigma factor, that is associated with the RNA polymerase and facilitates its binding to DNA. In eukaryotic cells, in contrast, promoter recognition is mediated by a suite of proteins called general **transcription factors** (**TFs**), which assemble at the promoter before, or simultaneously with, the binding of RNA polymerase II.

The assembly process starts with the binding of a general transcription factor to a short double-stranded DNA sequence primarily composed of T and A nucleotides. Because of its nucleotide makeup, this sequence is known as the **TATA box** [not to be confused with the −10 sequence of bacteria, which is similarly rich in T and A nucleotides (TATAAT)]. The TATA box is typically located 25 nucleotides upstream from the transcription start site. It is not the only DNA sequence that signals the start of transcription, but for many RNA polymerase II promoters it is the most important. The TATA box is recognized by a transcription factor called the **TATA-binding protein** (**TBP**). Interestingly, the TATA-binding protein binds in the minor groove, in contrast to most sequence-specific DNA binding proteins, as we discussed in Chapter 8. By binding in the minor groove, the TATA-binding protein induces a conspicuous kink in the helix, creating a physical landmark that highlights the location of the promoter (Figure 11).

The TATA-binding protein is only one of several transcription factors that assemble at the promoter and create what can be thought of as a landing pad for RNA polymerase II. In fact, the DNA-protein complex can be said to "recruit" the RNA polymerase in the sense that it creates a surface to which RNA polymerase binds via protein-protein interactions.

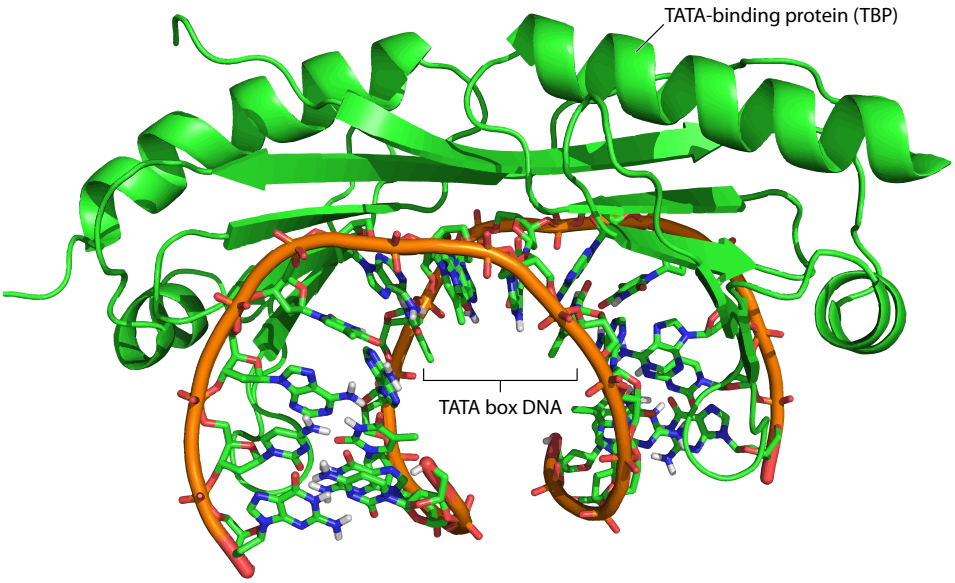

Figure 11 The binding of TATA-binding protein distorts the DNA double helix

TATA-binding protein (TBP) (*green*) binds to the specific DNA sequence found at the TATA box. Binding of TBP grossly distorts the shape of the DNA helix, as shown here.

Box 3 The appearance of the TATA-binding protein is an ancient branch point in the evolution of life

The TATA-binding protein is one of the most distinctive features of the eukaryotic transcription machinery, just as sigma factor is a hallmark of the bacterial machinery. Just how these two proteins, which represent two different modes of promoter recognition and are unrelated to each other, arose in evolution is a fascinating mystery. All contemporary life forms are believed to have arisen from a common ancestor, the Last Universal Common Ancestor (LUCA). This last common ancestor gave rise to Bacteria and a second branch from which contemporary Archaea and Eukaryotes evolved. Sigma factor is only found in the Bacteria branch, whereas the TATA-binding protein is featured in Archaea and Eukaryotes. Whereas Eukaryotes have many kinds of transcription factors, Archaea, which are more ancient and are believed to have given rise to Eukaryotes, have very few in addition to the TATA-binding protein. So the TATA-binding protein is believed to be a very ancient branch point in the evolution of life, and just how it and sigma factor arose is not known.

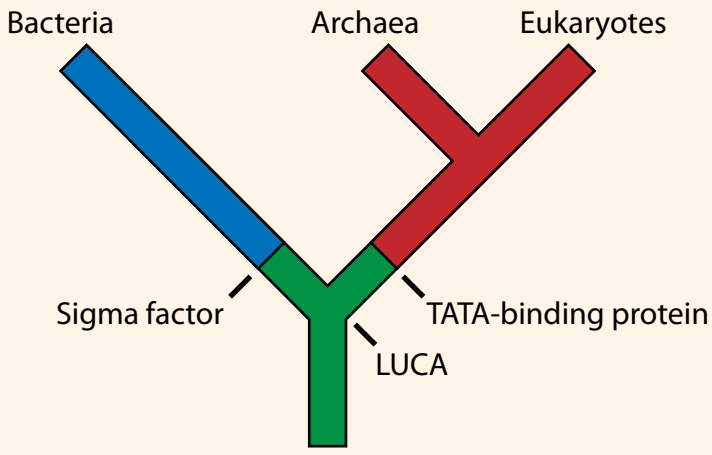

Figure 12 Sigma factor and TATA-binding protein arose after Archaea and Eukaryotes branched from Bacteria

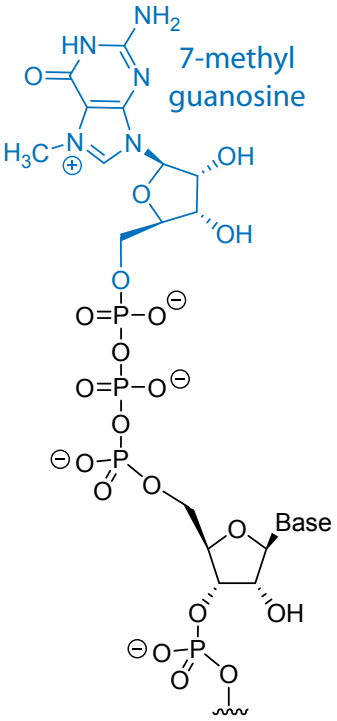

Figure 13 **RNA synthesized by RNA polymerase II is modified by the addition of a 5' cap**

RNA polymerase II-generated transcripts are covalently modified in three ways before they serve as messenger RNAs for protein synthesis

The differences between bacteria and eukaryotes in how messenger RNAs are generated does not end with the transcription machinery. Whereas nascent transcripts in bacterial cells are immediately ready to serve as messengers for protein synthesis, newly synthesized transcripts in the cells of higher organisms are in an immature, **pre-messenger RNA** (**pre-mRNA**) form, and these pre-mRNAs must be processed by three kinds of covalent modifications before they can exit the nucleus as mature messenger RNAs and serve as templates for protein synthesis.

Eukaryotic pre-mRNAs acquire a cap at their 5' terminus

The first modification occurs as soon as RNA polymerase II has produced about 25 nucleotides of RNA, at which point the 5' end of the new RNA molecule is modified by the addition of a modified guanosine nucleotide (Figure 13). What is remarkable about this guanine nucleotide is that it is attached to the 5' end of the nascent transcript by an unusual 5'-5' linkage via three phosphoryl groups. The guanine is additionally modified by the addition of a methyl group at the 7 position of the base. The entire structure comprising the reverse-linked guanosine nucleotide and the methylated base is known as a **cap**. This is an identifying feature of the 5' end of eukaryotic mRNAs and helps the cell distinguish mRNAs from the other types of RNAs present in the cell. RNA polymerases I and III produce RNAs without caps during transcription.

Eukaryotic transcripts acquire a poly-adenine nucleotide tail at their 3' terminus

A second covalent modification involving nucleotide addition takes place at the 3' end of the transcript, where a polymer of adenine nucleotides is attached. This **poly-A tail** is created in two steps. The first involves a cleavage reaction that cuts the transcript at the site where the tail will be added. Next, an enzyme known as **poly-A polymerase** adds approximately 200 adenine nucleotides to the 3' end generated by the cleavage (Figure 14). The nucleotide precursor for these additions is ATP, and the 5'-to-3'

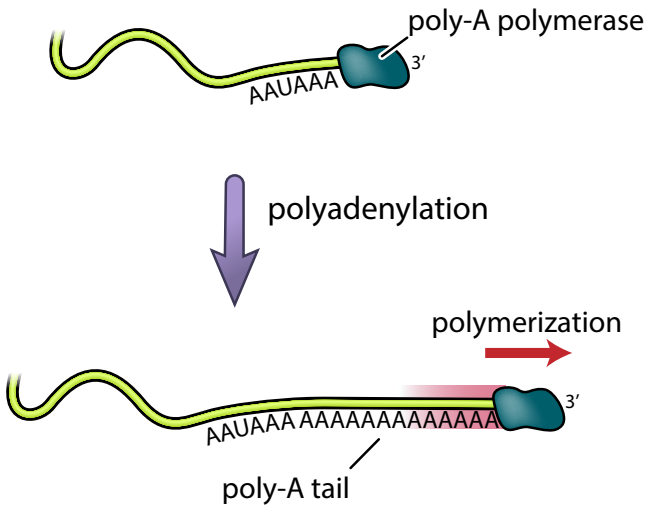

Figure 14 **A stretch of adenine nucleotides is added at the 3' end of the transcript by a poly-A polymerizing enzyme**

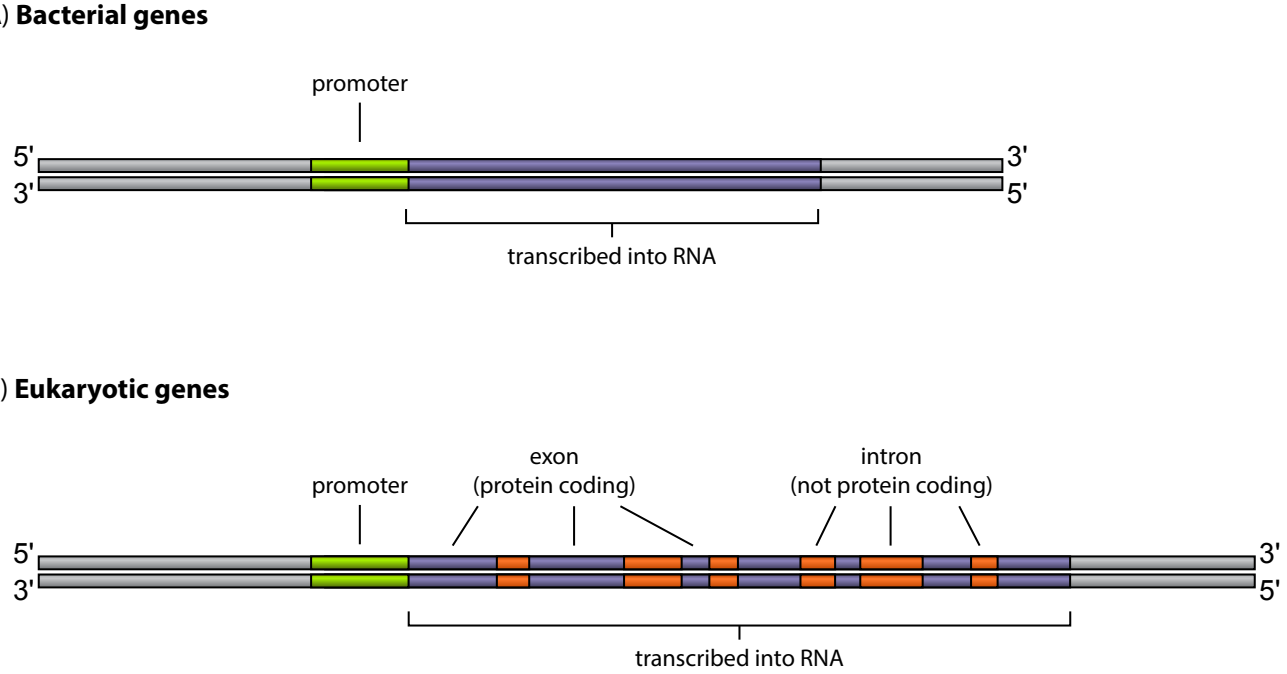

Figure 15 **Eukaryotic genes contain noncoding regions that are interspersed between protein-coding regions**

phosphodiester bonds are formed in the same way as during conventional RNA synthesis. Unlike other RNA polymerases, poly-A polymerase does not require a template; hence, the poly-A tails of eukaryotic mRNAs are not encoded in the genome.

Eukaryotic genes are interrupted by introns, which are removed from the pre-mRNA by splicing

The third and most spectacular modification of pre-mRNA stems from the fact that genes in the genomes of higher organisms are in pieces (Figure 15). That is, the protein-coding sequences are interrupted by non-coding stretches of DNA. The coding sequences are known as **exons** and the interruptions as **introns**. The number of introns in genes varies from as few as one to as many as hundreds and their size from less than 50 nucleotides to greater than a million nucleotides. In some cases the protein-coding portions constitute less than 10% the overall length of the gene. The transcription machinery does not discriminate between exons and introns, and the entire mosaic gene is copied into a single, long pre-mRNA. The introns are then removed after transcription by a process known as **splicing**.

Only after the pre-mRNA has acquired a cap and a tail and has had its introns removed is it ready to exit the nucleus and serve as an mRNA in the cytoplasm.

Introns are removed from pre-mRNA by two trans-esterification reactions

The removal of introns by splicing involves three positions in the pre-mRNA: the **5' splice junction**, the **3' splice junction**, and the **branch point adenosine** in the intron sequence (Figure 16). The two splice junctions

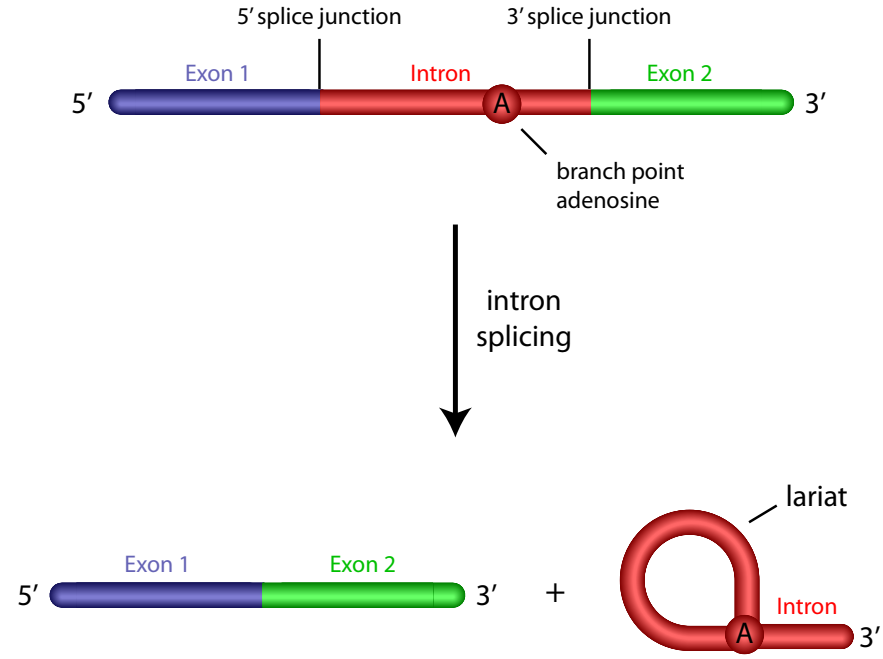

Figure 16 RNA splicing results in the removal of introns as lariat loops

mark the boundaries of the intron with the upstream and downstream exons, and the branch point adenosine is a particular adenosine nucleotide in the intron sequence. Each of these three sites has a consensus nucleotide sequence that is similar from intron to intron, providing the cell with cues as to where splicing is to take place.

Splicing involves two successive **transesterification** reactions in which one phosphodiester linkage is replaced by another. In the first reaction, the 2' hydroxyl group of the branch point adenosine attacks the phosphate group in the sugar-phosphate at the 5' splice junction. This cleaves the sugar-phosphate backbone of the RNA strand, creating a loop in the RNA molecule. This loop structure is known as a **lariat** (because it resembles the looped rope that cowboys use to lasso). Notice that the ribose of the branch point adenosine has three phosphodiester linkages; two at the 5' and 3' positions as part of the polynucleotide backbone and a newly created 2' linkage, which closes the loop to make the lariat. Lariat formation is an example of how the 2' hydroxyl enhances the versatility of RNA.

Formation of the lariat releases the 3' hydroxyl of the upstream (in the 5' direction) exon (exon 1 in Figure 17). In the second transesterification reaction, the released 3' hydroxyl attacks the phosphate at the 5' end of the downstream exon (exon 2), joining the two exons together and releasing the intron as a lariat. The two exons thereby become joined in a continuous coding sequence.

Splicing is catalyzed by a large complex of proteins and noncoding RNAs known as the **spliceosome**. The spliceosome is another example of a molecular machine. In this case the machine consists both of proteins and RNAs, and indeed the RNAs are central to its function, a point to which we return in Chapter 13. Consider that the spliceosome must carry out its task with exceptional accuracy. If the joining of two exons is inaccurate by even a single nucleotide, the resulting messenger RNA will have an extra nucleotide or will be missing a nucleotide. Because (as we will see in the next chapter) the ribosome translates coding sequences in successive units

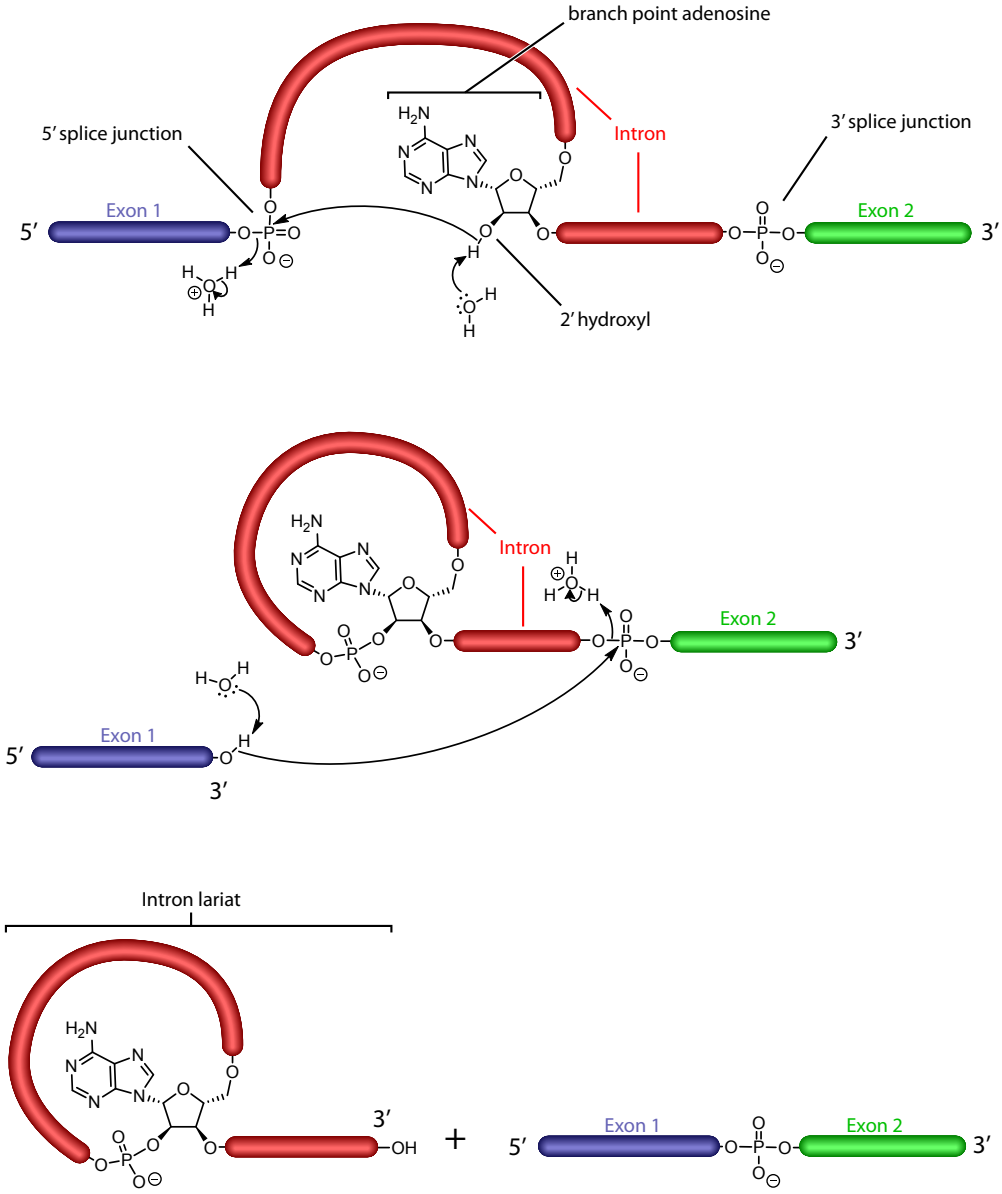

Figure 17 **RNA splicing takes place in two steps**

of three nucleotides for each amino acid, adding or subtracting a single nucleotide would shift the three-unit frame and result in a drastically altered protein-coding sequence. Therefore, the spliceosome needs to be nearly infallible in order to avoid causing catastrophic errors in translation.

Summary

RNA differs from DNA due to the presence of a 2' hydroxyl and the base uracil instead of thymine. Uracil and thymine pair with the same base, adenine.

RNA serves as an intermediary in the transmission of genetic information from the DNA double helix to the ribosome, the machine for protein synthesis. According to the central dogma, information in the form of the order of bases can be transmitted from one nucleic acid to another (as in DNA replication and transcription) and from nucleic acids to the order of amino acids in a protein (as in translation) but not from proteins to nucleic

acids.

Transcription is similar to replication in that the substrates are nucleotides and that DNA is used as a template for directing the order of nucleotide addition by base pairing. Transcription differs from replication in that, for any given transcription unit, only one strand of the double helix is copied (transcription is asymmetric) and that only limited stretches of the genome, namely gene-containing transcription units, are copied into RNA. RNA synthesis takes place in moving transcription bubbles in which the two strands of the DNA double helix separate from each other transiently while one serves as a template for polynucleotide synthesis. Nascent RNA briefly forms an RNA:DNA hybrid with the template strand and is then extruded as a free single strand. Either strand of the double helix can serve as a template, meaning that genes can be oriented in either direction on the genome.

The enzyme for transcription is RNA polymerase. Transcription commences at start sites, which are preceded by a promoter sequence. In bacteria, RNA polymerase consists of a core enzyme, which is responsible for RNA synthesis, and a sigma factor, which mediates promoter recognition and determines where RNA polymerase binds to DNA. The binding of RNA polymerase at the promoter causes the two strands of the double helix to separate, creating the transcription bubble, which then proceeds down the transcription unit as RNA synthesis proceeds.

Transcription in eukaryotes is more complex. Transcription of protein-coding genes is carried out by RNA polymerase II, one of three RNA polymerases. It recognizes promoters indirectly via binding to a complex of transcription factors that bind to the DNA and recruit the RNA polymerase. The most important of these factors is the TATA-binding protein, which binds to a promoter sequence known as the TATA box.

Transcription in eukaryotes is also more complex because the primary product of transcription, the pre-mRNA, undergoes three modifications before it serves as a messenger RNA for protein synthesis. It acquires a guanine nucleotide cap at its 5’ terminus, which is attached via an unusual 5’-to-5’ triphosphate linkage, and a poly-A tail at its 3’ terminus.

The most spectacular modification is splicing. Coding sequences in the genes of higher cells are interrupted by noncoding sequences known as introns. When these genes-in-pieces are copied into RNA, the introns must be removed by splicing so that the coding sequences, exons, can be joined together to form an uninterrupted messenger RNA. Splicing involves two transesterification reactions. In the first reaction, the 2’ hydroxyl of the branch point adenosine attacks the 3’ end of the upstream exon, releasing the exon and forming a lariat structure. In the next transesterification reaction, the newly created 3’ hydroxyl of the upstream exon attacks the phosphate at the junction of the intron and the downstream exon. This transesterification reaction releases the intron and fuses the upstream and downstream exons into a single coding sequence. Splicing is carried out by a molecular machine known as the spliceosome. The spliceosome is a complex of noncoding RNAs and proteins that mediates splicing with single-nucleotide precision.