# Transcription :

## *Reading the Instructions in the Genome*

*Using this book: This book is designed to be used in both introductory and advanced cell biology courses. The primary text is generally on the left side of the vertical divider, and printed in black. Details that are usually left to an advanced course are printed in blue and found on the right side of the divider. Finally, additional biomedically relevant information can be found in red print on either side of the divider.*

*Although DNA is an excellent medium for the storage of information, the very characteristic that makes it so stable and inherently self-correcting - being double-stranded - also makes it unwieldy for using that genetic information to make cell components. Since the informational parts of the molecule (the nitrogenous bases) are locked inside the ladder, reading it requires the energetically expensive task of breaking all the hydrogen bonds holding the two strands together. To do so for every single copy of each protein needed by the cell would not only take a lot of energy, but a lot of time. Instead, there must be a mechanism to take the information from DNA once (or a few times), and then make many copies of a protein from that single piece of information. That mechanism is transcription.*

In order to obtain the genetic information in a form that is easily read and then used to synthesize functioning proteins, the DNA must first be transcribed into RNA (ribonucleic acid). As we saw in chapter 1, RNA is extremely similar to DNA, using some of the same nitrogenous bases (adenine, guanine, cytosine) as well as one unique to RNA, uracil. Notice that uracil is very similar to thymine (chapter 7, fig.1), particularly in the placement and spacing of the hydrogen-bonding atoms. Since it is the hydrogen-bonding interaction of these bases (i.e. base-pairing of guanine to cytosine, adenine to thymine/uracil) that forms the basis of information transfer from original DNA to daughter cell DNA, it is logical to expect that the same kind of base-pairing mechanism is used to move the information from a storage state in the double-stranded nucleic acid (DNA) to a more useful/usable state in the form of a single-stranded nucleic acid (RNA).

The process of copying DNA into RNA is called transcription. In both prokaryotes and eukaryotes, transcription requires certain control elements (sequences of nucleotides within the DNA) to proceed properly. These elements are a promoter, a start site, and a stop site. The need for a recognizable point to begin and a point to end the process is fairly obvious. The promoter is somewhat different. The promoter controls the frequency of transcription. If you imagine the needs of a cell at any given time, clearly not all gene products are needed in the same quantity at the same time. There must be a way to control when or if transcription occurs, and at what speed.

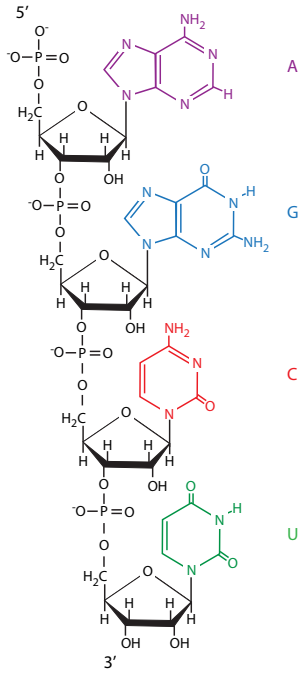


Figure 1. Ribonucleic Acid (RNA) is a polymer of the nucleotides adenine, guanine, cytosine, and uracil, connected by 3'-5' phosphodiester bonds.

The bare-bones version of the process goes something like this: (1) special docking proteins recognize the promoter sequence and bind to it, unzipping a small section around the "start" site; (2) RNA polymerase binds to those special proteins and to the little bit of single-stranded DNA that has just opened up; (3) a helicase enzyme (part of, or attached to the polymerase) unzips the DNA; (4) the RNA polymerase follows behind the helicase, "reading" the DNA sequence, taking ribonucleotides from the environment, matching them against the DNA template, and if they match, adding them to the previous ribonucleotide or RNA chain. This continues until the polymerase reaches the stop site, at which point, it detaches from the template DNA, also releasing the newly made RNA copy of that DNA. Of course, if that was all there was to it, there wouldn't be entire journals dedicated to studying RNA, its transcription, and the control of that transcription.

The sequence of the promoter is directly related to its function. There may be promoters for housekeeping genes (needed constantly, but at low copy number), "normal" genes (needed as the cell's situation dictates, rate of transcription also varies), stress response genes (needed rarely), and a variety of other categories. Even within a category, the sequence of the promoter determines its strength. This is based upon what is known as the "consensus sequence". The consensus sequence is a theoretical "best" promoter based on a survey of all genes in a particular category. The figure below shows an alignment of the promoter sequences of a variety of different genes, all of which are regulated by the same type of promoter and promoter-binding-protein. The highlighted boxes show areas centered around -35 (35 nucleotides upstream of the start site) and -10.

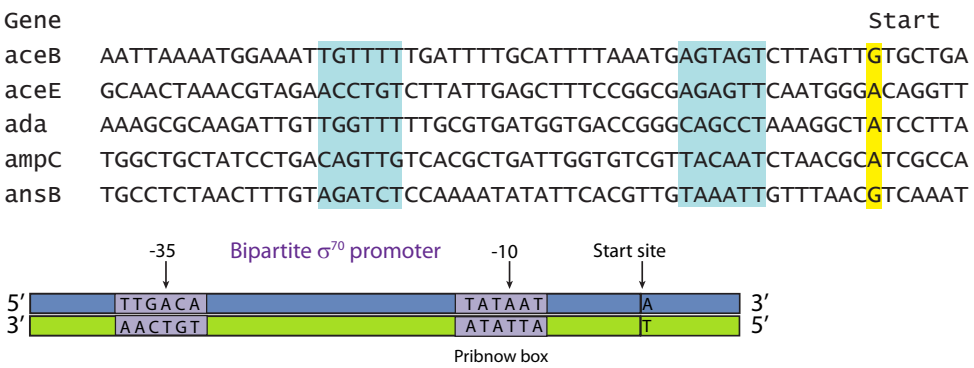


Figure 2. Upstream control sequences of various E. coli genes. Below it is the consensus sequence for for $\sigma^{70}$ promoters.

In contrast to its cellular role as a transient and disposable carrier of genetic information, RNA is thought to have been the primary molecule responsible for making life possible on earth. It has long been postulated that it served dual roles as both a repository of genetic information and as a rudimentary enzyme to act upon that information. Unfortunately, prebiotic chemists have been stymied for decades in coming up with a reasonable synthetic pathway by which RNA could arise from the simple molecules of the earth's primordial "soup". The key problem was that ribose could be synthesized, though not particularly efficiently, and bases could be synthesized, but there was no way to connect them together. The chemistry would not allow a condensation reaction between the bases and sugars. In 2009, by leaving behind the conventional idea that ribonucleotides must have been synthesized from ribose and purines/pyrimidines, Powner, Gerland, and Sutherland (*Nature* 459:239-242, 2009) showed that in fact, ribonucleotides could be synthesized from the chemical conditions of a newly formed earth. Rather than attempt to make each "part" and put them together, Powner et al synthesized a molecule that contained parts of what would eventually be both the ribose and a pyrimidine, 2-aminooxazole. Through a series of reactions that utilized phosphate as a catalyst and scavenger, all of which were clearly plausible in the current model of primordial earth, both ribocytidine and ribouridine were created. Of course, this is only the beginning, since this does not extend directly to the formation of purine nucleotides, but it is a very significant step in prebiotic chemistry, and an excellent example of the virtues of "stepping outside of the box" sometimes.

The consensus sequence in fig. 2 shows the most common nucleotide found at each position within those areas of similarity. In this example, the most common prokaryotic promoter is shown: the $\sigma^{70}$ promoter, so called because it is recognized and bound by the $\sigma^{70}$ transcription factor. [Here, and by universally accepted convention, recognition sequences of DNA are written as the nucleotides would occur from 5' to 3' on the sense, or non-template, strand. ] It is a two-part promoter, with a region centered around -35 (consensus TTGACA), and a region (sometimes called Pribnow box, consensus TATAAT) centered around -10. The (-) sign indicates that the nucleotide is "upstream" of the start site. Upstream means "to the left" when the nucleotides are written as a string of letters, and it means "on the 5' side of" with respect to the 5'-3' directionality of a DNA strand. Notice the relationship between the various individual promoters and the consensus sequence. In general, those promoters with more matches to the consensus sequence are stronger promoters.

A few paragraphs ago, the task of the promoter was defined as controlling the frequency of transcription. How does it do that? What does it mean to be a stronger (or weaker) promoter? First, keep in mind that the expression of any given gene is not automatic, or 100%. At any point in time, many of a cell's genes will be near 0%, or shut off. However, even genes that are turned on are transcribed at different rates. One of the governing factors is the recognition of the promoter site by the RNA Polymerase. For stronger promoters, the RNA polymerase is more likely to recognize the site, dock properly, open up the double helix, and begin transcribing. On the other hand, the RNA polymerase can potentially recognize weaker promoters, but it is less likely to do so, instead passing it by as just another unimportant stretch of DNA. While this is partially a matter of recognition by the polymerase, keep in mind that it is actually governed by recognition of the promoter sequence by the general transcription factors (to be discussed shortly) such as sigma factors in prokaryotes that are recognized by the polymerase.

Notice that there is a high proportion of (A)denines and (T)hymines in the $\sigma^{70}$ promoter sequences. This is true for many promoters in both prokaryotic and eukaryotic genes. As you probably suspected, this is advantageous because there are only two H-bonds between A-T pairs (as opposed to 3 H-bonds between G-C pairs), which means that it is 33% easier to unzip.

In E. coli, as with other prokaryotes, there is only one true RNA polymerase (not including the specialty RNA polymerase, primase, which makes short RNA primers for DNA replication). The polymerase is a multi-subunit holoenzyme comprised primarily of two α subunits, a β subunit, a β' subunit, an ω subunit, and a σ subunit. The α subunits are primarily structural, assembling the holoenzyme and associated regulatory factors. The β subunit contains the polymerase activity that catalyzes the synthesis of RNA, while the β' subunit is used to nonspecifically bind to DNA. The ω subunit is involved in assembly of the holoenzyme and may also play a role in maintaining the structural integrity of the RNA polymerase. Finally, there is the σ subunit, which does not stay closely associated with the core enzyme (αββ'ω) except when helping to initiate transcription, and is used to recognize the promoter by simultaneously decreasing the affinity of RNAP to DNA in general, but increasing the affinity of RNAP for specific DNA promoter sequences. Why decrease the affinity for non-specific DNA? When the RNAP is not in use, it does not just float about in the nucleoplasm: it is bound quite tightly along the DNA. When the sigma is bound, the decreased affinity allows the RNAP holoenzyme to move along the DNA and scan for promoter sequences. There are multiple isoforms of the σ subunit (such as the sigma-70 mentioned above), each of which recognizes different promoter sequences. All isoforms perform the same basic function of properly locating the RNAP to the start of a gene, and all isoforms only stay attached to the holoenzyme for that one transient purpose, after which they are released (usually after transcribing about ten nucleotides).

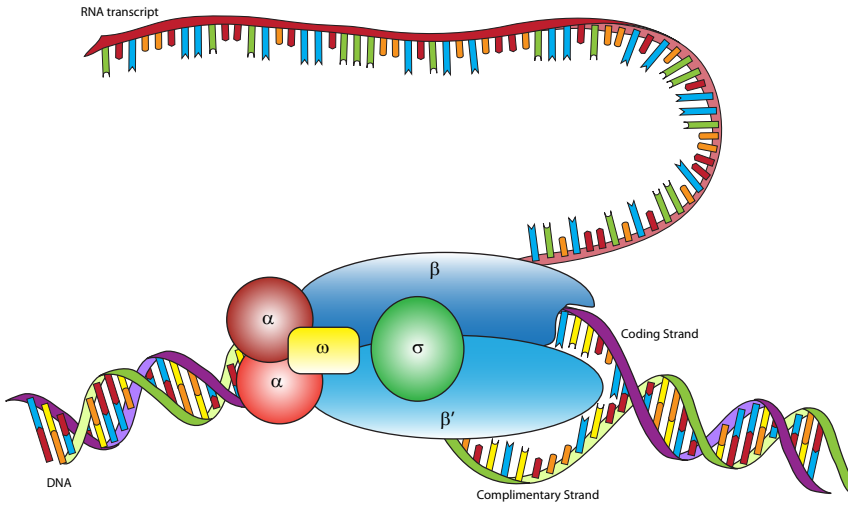


Figure 3. Prokaryotic RNA Polymerases consist of two α subunits, a β, β', ω, and σ subunits.

Although RNA polymerase was discovered in 1960, the E. coli RNAP has not yet been successfully mapped by x-ray crystallography. However, it is very similar to the RNAP of the archaean species, Thermophilus aquaticus, which is highly stable (= easier to crystallize) and for which an x-ray crystallographic structure has been elucidated. The data from the Taq RNAP structure and electron microscopic analyses of E. coli RNAP produce a lobster-claw-shaped holoenzyme. The inner surface of the claw is lined with positively charged amino acids that can interact with the negatively charged DNA, and when the holoenzyme binds a sigma subunit, the two halves of the claw (formed mostly by the beta and beta' subunits) move closer together to interact with the DNA.

Rifamycins are a class of antibiotics that include rifamycin B, made by the bacteria Streptomyces mediterranei (incidentally just one of many antibiotics derived from the Streptomyces genus), and rifampicin, its synthetic cousin. They work by binding within the DNA-RNA channel near the active site of RNA polymerase, which sterically prevents the addition of nucleotides to the RNA strand. If the organism cannot transcribe RNA, it cannot use the RNA to make the enzymes and other proteins necessary for life either, and dies. The rifamycin binding site is highly conserved in most prokaryotes but not in eukaryotes, so the antibiotic kills bacteria specifically with little chance of harm to eukaryotes.

Once the holoenzyme has recognized and bound tightly to the DNA at the promoter site, the next step is to "melt" the DNA (breaking the H-bonds and separating the strands of the double helix) in that area so that the RNAP can proceed downstream, read the template DNA strand, and produce the new RNA. Often many RNA transcripts of a gene are needed to produce a large number of active proteins in a short span of time. Highly transcriptionally active genes therefore often have multiple RNA polymerases reading them, one right after another. Generally, an RNA polymerase only needs to process about 15 nucleotides before there is room for another RNAP can bind the promoter and start another transcript.

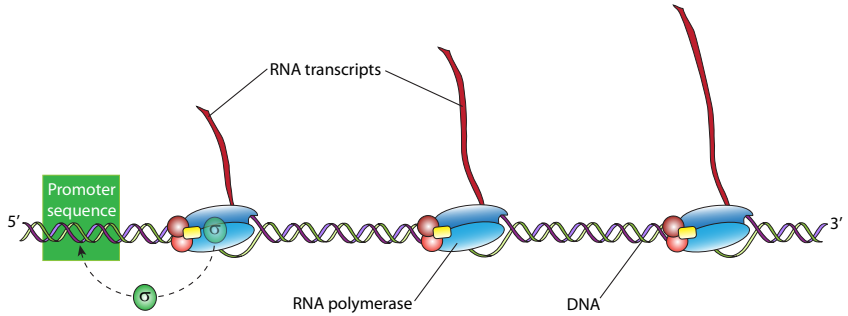


Figure 4. A new RNA polymerase can begin transcribing a gene before the previous one has finished.

The elongation phase of transcription proceeds in a 5' to 3' direction, which is to say that new nucleotides are added to the 3'-OH of the growing strand. Elongation is a stochastic process in which one of the plentiful free-floating ribonucleotides drops into the active site of RNAP opposite the DNA template. If it is the correct nucleotide (complementary to the template), then H-bonds will temporarily form, stabilizing the new nucleotide in place long enough for the RNAP to catalyze the formation of a phosphodiester bond between the 3'-OH of the RNA-in-progress and the 5'-phosphate of the nucleotide. However, if it is the incorrect nucleotide, the proper H-bonds do not form, and the nucleotide usually dissociates from the active site before the RNAP has a chance to bind it to the growing RNA strand. Obviously, this is not a perfect system, and in fact, the error rate for transcription is quite high at approximately 1 in 10000 nucleotides. Fortunately, the cell generally churns out many copies of RNA from any given gene very quickly (approximately 80 nucleotides per second), most of which are either error-free or have errors that do not affect the function of the end-product protein. Furthermore, unlike DNA, in which errors of replication get carried along from one generation of cells to the next, RNA is not a storage medium, and its transient nature means that even mutations that severely impact the protein function only affect the few proteins translated from that one RNA, not the proteins generated from other RNAs made from the same template gene, much less subsequent generations. In other words, to misappropriate a phrase from the movie Meatballs, "It just doesn't matter."

Strand separation is an energetically difficult process due to the strength of the combined H-bonds between the strands, and often an RNAP may make several short-lived abortive attempts before finally prying open the double helix long and far enough to allow the RNAP to stabilize and transcribe continuously to the stop site.

Eventually, the RNA polymerase reaches the end of the gene and stops transcribing. The termination site is usually marked by a sequence of 4-10 adenine (A) residues on the template strand, and some have a palindromic G-C-rich region that forms a hairpin loop just upstream of the series of adenines. In the first case, it is thought that the resulting string of A-U base pairs is unstable and may lead to the RNAP and the new RNA strand falling off the template DNA while at the same time, the hairpin structure may cause the RNAP to stop or pause, and this can also lead to it dissociating from the DNA. Only about half of all transcription termination sites are marked in this way though, and the others have no significant hairpin loops or easily recognizable sequences other than a series of G-C-rich regions. In this type of termination site, the enzymatic co-factor, rho, is required for termination, and so this is known as rho-dependent termination. Rho is an RNA-binding protein with helicase activity, so it is postulated that it effects termination by forcing the RNA strand off of the DNA template.

## *Eukaryotic Transcription*

Transcription in eukaryotes is more complicated, but follows the same general ideas. The promoter sequences are much more varied both in placement (with respect to the start site) and size. As we will see in the next chapter, eukaryotic genes have many more control elements regulating their expression than do prokaryotic genes. Not only are there more control elements, there are also more RNA polymerases, which serve different specific cellular functions. Obviously, the broad function and location of all the RNA polymerases is the same: read a DNA template and transcribe an RNA copy of it; and since the DNA is found only in the nucleus, so are the polymerases. However, the polymerases differ in exactly what kinds of RNA they produce. RNA Polymerase I is specialized for producing pre-rRNA (rRNA = ribosomal RNA). The pre-rRNA is cleaved post-transcriptionally and incorporated into the ribosomes. Since ribosomes are assembled in the the nucleolus, that is the part of the nucleus in which most RNA Polymerase I is concentrated. RNA Polymerase III also makes an RNA (5S) that is incorporated into the ribosome. It also makes other *untranslated* RNAs such as tRNAs and a variety of small nuclear RNAs. The only RNA polymerase that makes the translatable RNA (mRNA, or messenger RNA) that most people think of when RNA is referred to generically, is RNA polymerase II. This is the RNA polymerase that produces pre-mRNA, which after some processing, becomes mRNA, is transported out of the nucleus, and finally translated into proteins. All of the eukaryotic RNA polymerases are composed of two large subunits, roughly analogous to the β and β' subunits of prokaryotic RNAP, but instead of just three or four other subunits, there are over a dozen smaller subunits to the eukaryotic RNA polymerase holoenzymes.

Initiation of transcription is also much more complicated. Not only is there great variety in promoters recognized by RNAP II, both RNAP I and RNAP III recognize promoters with particular structural characteristics. One of the most common eukaryotic RNAP II promoters is the TATA box, named for the highly conserved motif that defines it. Although it appears similar to the Pribnow box in prokaryotes, it is generally located further upstream from the start site, and its position is far more variable. Whereas the Pribnow box is located at -10, the TATA box may be located closer to -30 +/- 4. Also, rather than just a sigma factor to recognize the promoter in conjunction with the polymerase core enzyme, the eukaryotic promoter is recognized by a multi-subunit complex called transcription factor IID (TFIID). TFIID is comprised of TATA-binding protein (TBP)

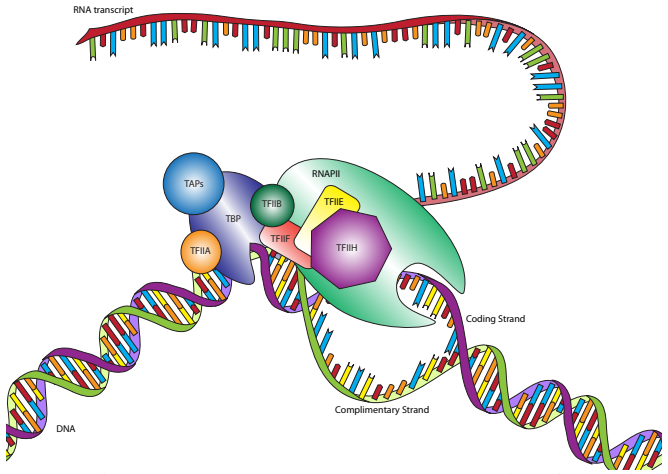


Figure 5. Eukaryotic Transcription. An initiation complex of several transcription factors is needed to dock the RNA Polymerase II in position to begin transcription.

and several TBP-associated factors (TAFs). This binding of the promoter by TFIID occurs independently of RNA Polymerase II, and in fact, RNAP II will not attach to TFIID at this time. After TFIID has bound the TATA box, two more transcription factors, TFIIA and TFIIB, attach to the TFIID as well as the nearby DNA, stabilizing the complex. TFIIF attaches to TFIID and TFIIB to allow docking of the RNA Polymerase II. The complex is still not ready to begin transcription: two more factors are required. TFIIE binds TFIIF and RNAP II, and finally, TFIIH attaches to RNAP II, providing a helicase activity needed to pry apart the two strands of DNA and allow the polymerase to read one of them. TFIIH also has another important enzymatic activity: it is also a serine kinase that phosphorylates the carboxyl-terminal domain (CTD) of RNA polymerase II. There are several serines in the CTD, and as they are sequentially phosphorylated, the CTD extends like a (negatively charged) tail and helps to promote separation between the RNAP II and the TFIID/promoter.

The eukaryotic RNA polymerases were named I, II, and III based on their elution order from ion-exchange chromatography purification. They are also partially distinguishable by their sensitivity to $\alpha$-amanitin and related amatoxin-family mushroom poisons. RNAP I (and prokaryotic RNAP) is insensitive to these toxins, RNAP III is somewhat sensitive ($K_d$ ~$10^{-6}$ M), and RNAP II is highly sensitive ($K_d$ ~$10^{-8}$ M). These toxins act by binding to a site in the RNA-DNA cleft and interfering with translocation of the RNA. That is, there is no problem with importing a nucleotide or with attaching it to the new RNA, but the RNA strand cannot move through the active site and allow the next nucleotide to be added.

Elongation of the RNA strand in eukaryotes is very similar to that in prokaryotes with the obvious difference that transcription occurs in the nucleus rather than in the cytoplasm. Thus, in prokaryotes, the RNA can be used for translation of proteins even as it is still being transcribed from the DNA! In eukaryotes, the situation is significantly more complex: there are a number of post-transcriptional events (5' end-capping, 3' polyadenylation, and often RNA splicing) that must occur before the RNA is ready to be transported out of the nucleus and made available for translation in the cytoplasm.

Termination of eukaryotic transcription is not well-described at this writing. RNAP I appears to require a DNA-binding termination factor, which is not analogous to the prokaryotic Rho factor, which is an RNA binding protein. RNAP III terminates transcription without any external factor, and this termination usually occurs after adding a series of uridine residues. However, it does not appear to use the hairpin loop structure found in rho-independent bacterial transcription. The termination of protein-coding RNAP II transcripts is linked to an enzyme complex that also cleaves part of the 3' end of the RNA off, and adds a poly-A tail. However, it is not clear how the polyadenylation complex is involved in determining the point of transcription termination, which can be over 1000 nucleotides beyond the poly-A site (e.g. the β-globin gene in Mus musculus). Upon termination and release from the RNAP II and template DNA, the RNA is known as the primary transcript, but must undergo post-transcriptional processing before it is a mature messenger RNA (mRNA) ready to be exported to the cytoplasm and used to direct translation.

## *Post-Transcriptional Processing of RNA*

The first of the post-transcriptional events is 5' end capping. Once the 5' end of a nascent RNA extends free of the RNAP II approximately 20-30 nt, it is ready to be capped by a 7-methylguanosine structure. This 5' "cap" serves as a recognition site for transport of the completed mRNA out of the nucleus and into the cytoplasm.

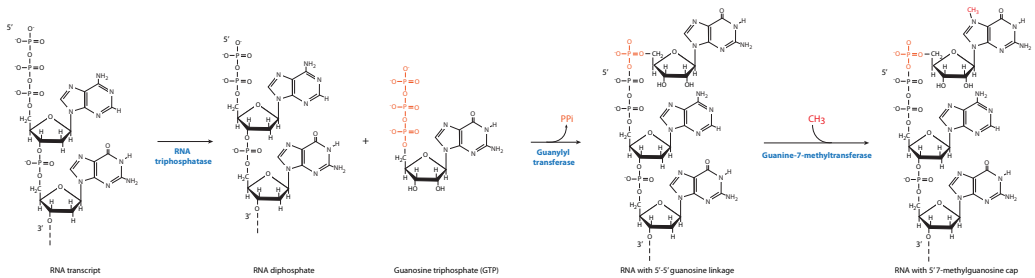


Figure 6. Capping the 5' end of eukaryotic transcripts.

The process actually involves three steps. First, RNA triphosphatase removes the 5'-terminal triphosphate group. Guanylation by GTP is catalyzed by capping enzyme, forming an unusual 5'-5' "backward" bond between the new guanine and the first nucleotide of the RNA transcript. Finally, guanine-7-methyltransferase methylates the newly attached guanine.

On the opposite end of the RNA, on the free 3'-OH, *polyadenylation* occurs. As noted previously, an enzyme complex that docks to a site on the CTD tail of RNAP II cleaves a portion of the 3' end near an AAUAAA recognition sequence and then serially adds a large number of adenine residues. The poly(A) tail is not required for translation, but it has an effect on the stability of transcripts in the cytoplasm. As mRNA molecules stay in the cytoplasm longer, the poly(A) tail is gradually removed. Once the poly(A) tail is gone, the mRNA will soon be destroyed. mRNA molecules with longer poly(A) tails are generally longer-lived in the cytoplasm than those with shorter tails, but there is currently no evidence for a directly proportional effect.

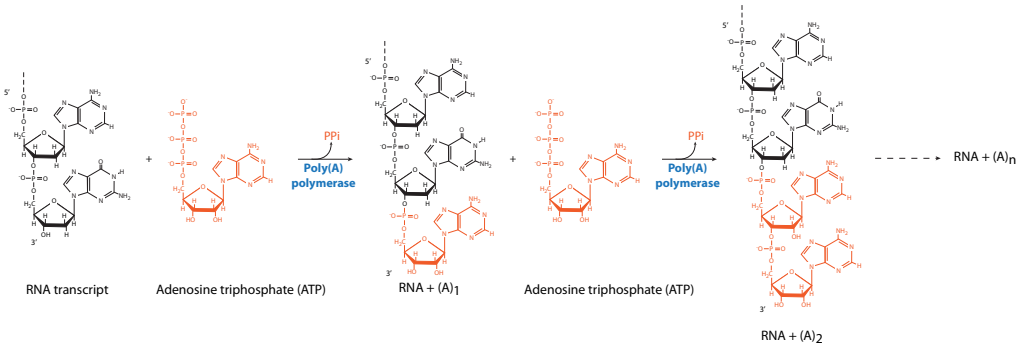


Figure 7. Polyadenylation of the primary transcript.

The third and most complicated modification to newly-transcribed eukaryotic RNA is *splicing*. Unlike prokaryotic RNA, which is a continuously translatable coding region immediately as it comes out of the RNA polymerase, most eukaryotic RNAs have interrupted coding regions. Splicing is the process by which the non-coding regions, known as *introns*, are removed, and the coding regions, known as *exons*, are connected together. In some RNAs, this can happen autonomously, with part of the RNA acting as an enzymatic catalyst for the process. This requires that the RNA have a specific secondary and tertiary structure, bringing the two exons close together while looping out the intron. It was the study of this phenomenon that led to the discovery of ribozymes, which are enzymes made of RNA.

In most cases, however, splicing is carried out by a multi-subunit protein complex known as the spliceosome. Whether it is self-spliced or by spliceosome, there are three main sequence components needed to define an intron that is going to be spliced out (fig. 8). There is a *5' splice site* with the consensus sequence AG|GUAAGU. There is a *3' splice site* that starts with an 11-nucleotide polypyrimidine tract followed by NCAG|G. And somewhere in between the two, there is a branchpoint adenine, typically within a YNCURAY sequence (Y is a pyrimidine, N is any nucleotide, R is a purine). Splicing is actually a set of two sequential transesterification reactions, and requires physical prox-

Although the enzyme that cleaves the primary transcript in preparation for polyadenylation has not been identified, two nonenzymatic factors, the excitingly-named cleavage factor I (CFI) and cleavage factor II (CFII) have been implicated. The serial adenylation comes from the activity of poly(A) polymerase (PAP) in conjunction with CPSF (cleavage and polyadenylation specificity factor), which binds to the RNA. PAP itself has relatively poor affinity for RNA. As with other nucleic acid polymerases, it adds new nucleotides onto the free 3'-OH of the pre-existing chain. To encourage processivity (continuous polymerization) poly(A) binding protein II (PABII) joins the polyadenylation complex, and is involved in controlling the final length of the poly(A) tail. It should be noted that PABII is a nuclear protein and should not be confused with PABP (poly(A) binding protein) which binds to mRNA molecules in the cytoplasm and plays a role in protecting them from nuclease attack.

Until the discovery of ribozymes, it had been assumed that only the enzymes could only be generated with the diversity of structures possible with the amino acids in proteins.

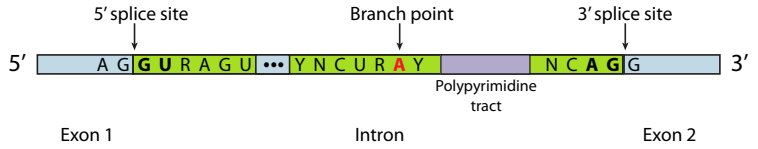


Figure 8. Consensus sequences for splicing.

imity of the reactive sites by bending and looping of the RNA, either autonomously or around protein factors know as snRNPs (pronounced "snurps"). SnRNPs is an acronym for small nuclear ribonucleoproteins. They contain both a protein and a small nuclear RNA (snRNA) component; the latter helps with sequence recognition. Examination of the structure of the snRNA part of these spliceosome snRNPs shows that they are very similar to the shapes taken by the RNA transcript itself in cases of self-splicing. Keeping that in mind, much of the following description of spliceosome-mediated splicing happens in self-splicing as well.

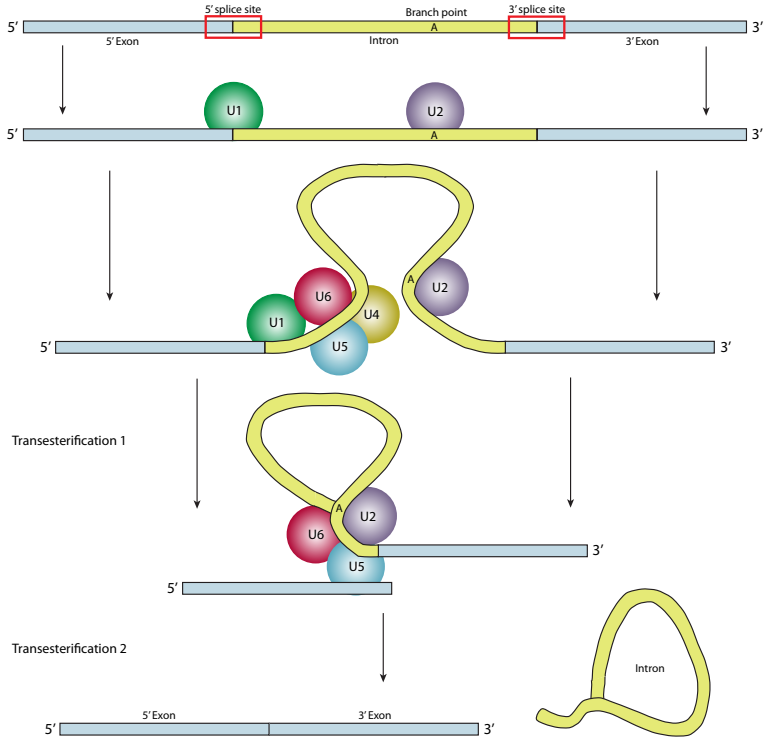


Figure 9. RNA splicing by spliceosome. Description in text below.

In the first step, the U1 snRNP binds to the intronic portion of the 5' splice site. Next, the U2 snRNP binds to the consensus site around the branchpoint, but importantly, there is no base-pairing to the branchpoint A itself. Instead, due to basepairing of U2 with the surrounding sequence, the branchpoint A is forced to bulge out from the rest of the RNA in that region. U4, U5, and U6 join the spliceosome together, but while U5 binds to the 5' exon, and U6 displaces U1 at the 5' splice site, U4 is only transiently attached and also falls off the spliceosome before the first transesterification reaction. As the figure shows, in this reaction, the 2'-OH of the branchpoint A nucleophilically at-

Although the snRNPs are the primary components of the spliceosome, a variety of other splicing factors also play a role. The most prominent are U2AF (U2-associated factor, which binds to the polypyrimidine tract, and SF1 (splicing factor 1, aka branchpoint protein BPP) which binds to consensus sequence near the branchpoint. Together they help to properly position the U2 snRNP. There are also a variety of other less-studied splicing factors from the SR protein family (C-terminal Serine-Arginine binding motif) and the hnRNP (heterogenous nuclear ribonucleoprotein) families that act to recruit the primary members of the spliceosome to their proper locations.

tacks the 5'-phosphate of the first intron nucleotide to form a lariat structure in which the 5' end of the intron is connected to the branchpoint via a 2',5'-phosphodiester bond. This releases the 5' exon (and the whole 5' half of the RNA for that matter), but it is kept in close proximity to the 3' exon (and the rest of the RNA) by U5, which attaches to both exons. This allows the second transesterification to take place, in which the 3'-OH of the first exon attacks the 5' phosphate at the beginning of the second exon, thus simultaneously breaking the bond between the intron and the second exon, and also connecting the two exons via a conventional 3',5'-phosphodiester bond. The intron, in the shape of a lariat, is thus released and will be quickly degraded.

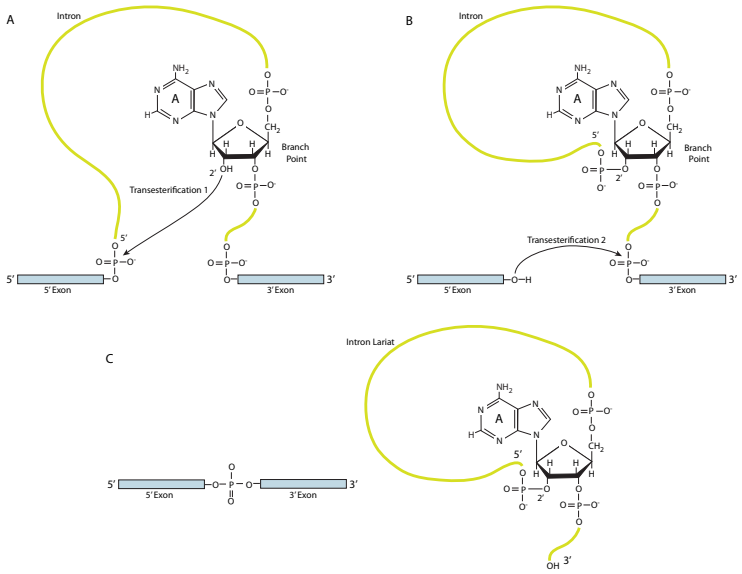


Figure 10. (A) Transesterification reaction 1 connects the 2'-OH of the branchpoint ribose to the 5'-phosphate of the 3' end of exon 1. (B) Transesterification 2 is an attack by the remaining -OH on the 3' end of exon 1 on the 5' phosphate of exon 2. (C) This simultaneously releases the lariat and connects the exons.

Splicing is an efficient (with respect to genome size) way to generate protein diversity. In alternative splicing, some potential introns may be spliced out under certain circumstances but remain as coding sequence under other circumstances. Recall that the splice sites are recognized by base-pairing and therefore, there can be stronger and weaker splice sites depending on how close they are to the consensus and the complementary sequence on the snRNPs. Therefore, a gene with several potential introns may have all introns spliced out 80% of the time, but the other 20% of the time, perhaps only one or two introns are spliced out. Adding variability, there are splicing factors that may bind near splice sites and can either make them more easily recognizable, or nearly hidden.

The classic example of alternative splicing is the gene encoding α-tropomyosin (fig. 11). By splicing in/out different combinations of exons, a single gene can generate seven different proteins, depending on the tissue type. In these cases, particular types of cells or tissues contain specific combinations of splicing factors, and therefore control the recognition of specific splice sites, leading to the different splicing patterns.

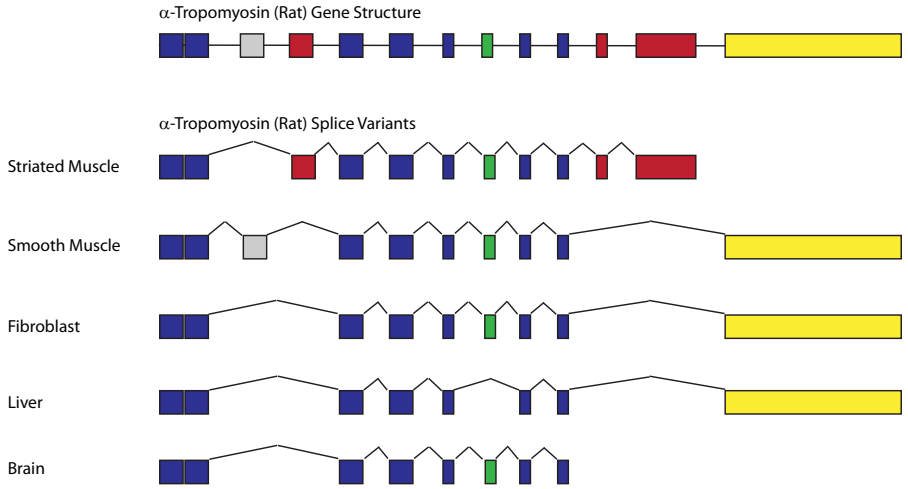


Figure 11. Alternative splicing of the a-tropomyosin gene leads to different forms of the mRNA and protein in different cell types.

Although this concludes the discussion of basic mechanisms of transcription, the next chapter is really a continuation of this one: control of gene expression in its simplest form is regulating the recognition of a promoter sequence by an RNA polymerase.

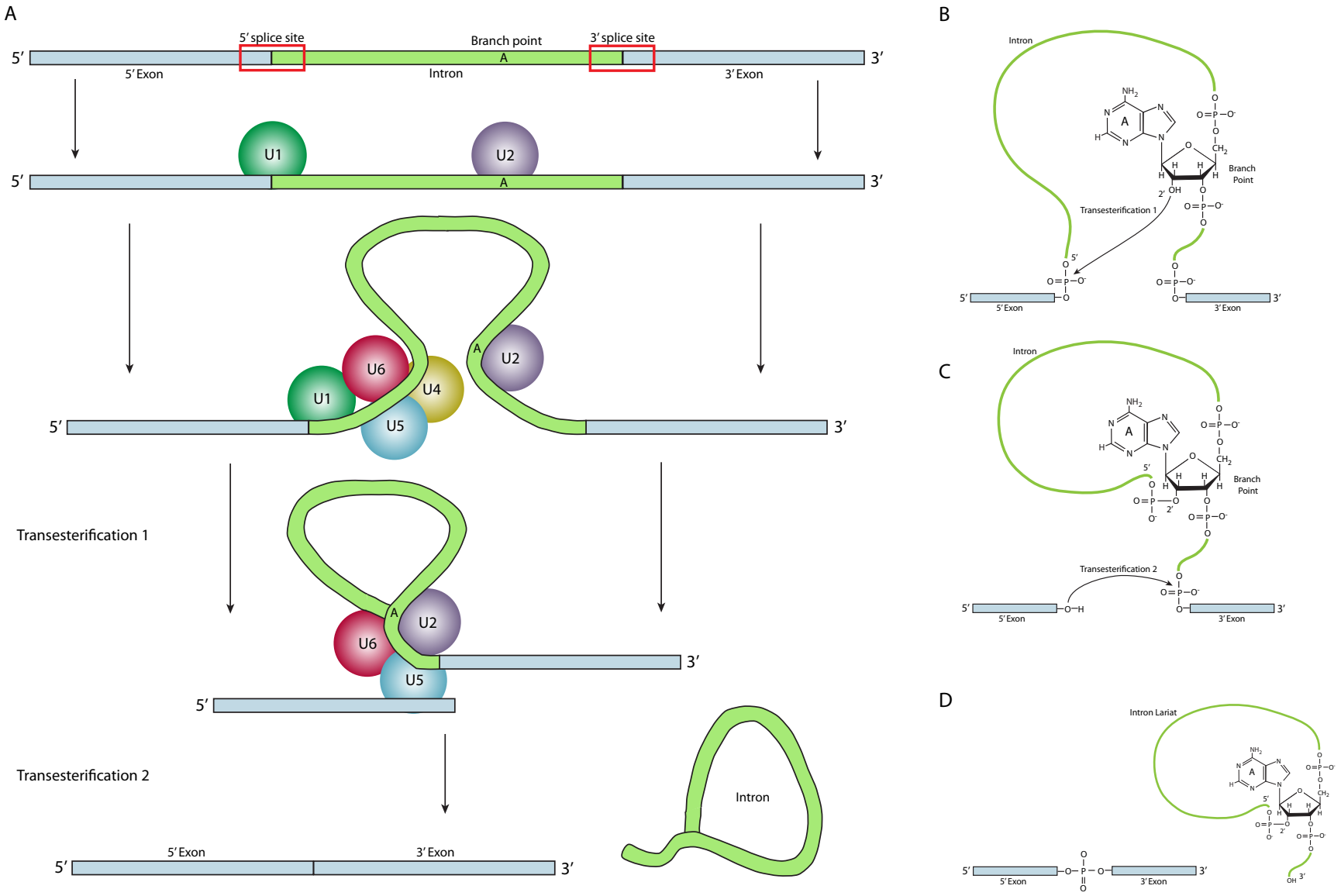


Figure 12. Splicing of RNA. Splicing removes some portions of a primary transcript (introns) while combining the remaining RNA (exons) to form the final mRNA sequence. The chemical mechanism is a series of two transesterifications. The first accomplishes the looping of the intron, while the second releases that intron as a "lariat" while simultaneously joining the two exons together. Splicing specificity relies on three landmarks on the RNA, most important of which is the branchpoint, an adenine residue that is the connection point for the intron loop formation. The other two landmarks are the 3' and the 5' splice sites, each of which are sequences that bind to snRNPs to bring together the spliceosome.