INTEGRATED DNA TECHNOLOGIES

Protein Synthesis

Contents

Introduction

The genetic message carried on the DNA molecule is a code. This code is ultimately translated into a sequence of amino acids that, when complete, becomes a protein. Proteins carry out the "business" of the cell. Some proteins are used as structural components of cells, some are used to transport other molecules, still others are charged with directing chemical reactions. The latter class of proteins is the enzymes. Regardless of the role played by a protein in the cell one aspect is the same, they are all encoded in the base sequences of DNA. The path from DNA sequence to protein sequence is an elegant but complex process that is composed of two major steps. The first is **transcription**, in which DNA is converted into a mature messenger RNA (mRNA), and the second is **translation**, in which the base sequence of the mRNA is "read" and converted into an amino acid sequence.

DNA Transcription

A gene is composed of a specific sequence of bases but only some of these bases actually carry the genetic message for making a protein. The rest of the DNA bases in a gene are divided among a wide range of functions, only some of which are fully understood. It is possible to describe what could be called a "generic gene" but it must be remembered that for every "average" feature of such a gene there will be hundreds, if not thousands, of exceptions, modifications, and special cases. With this in mind, the anatomy of a generic gene found in animals and plants is shown in Figure 1. The first convention used to represent genes is that the DNA sequences used to code for protein are shown as boxes and the DNA sequences that are part of the gene but do not code for protein are shown as lines.

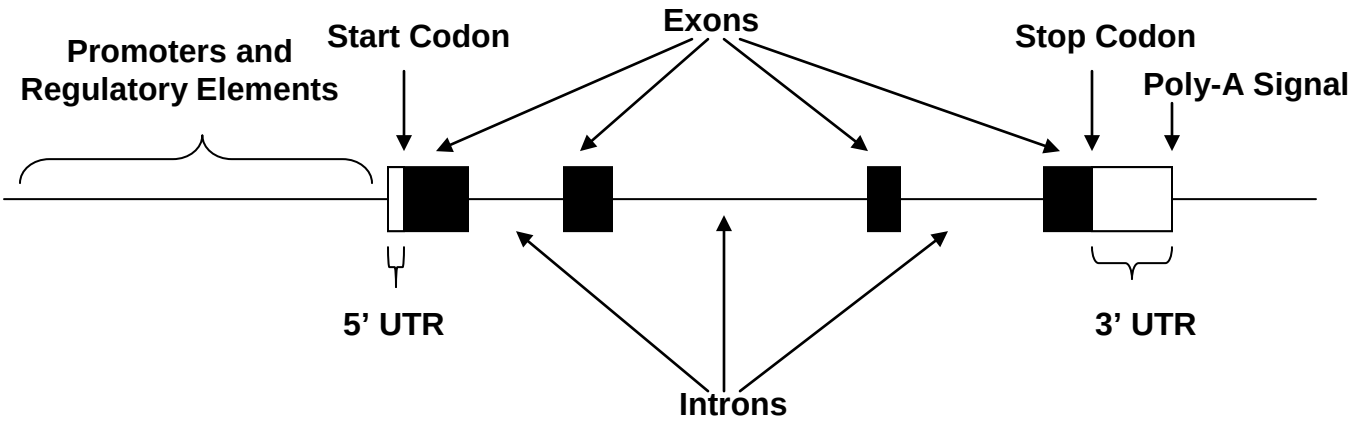


Figure 1. Anatomy of a eukaryotic gene.

The first feature of this generic gene that should be mentioned is that the part of the gene that codes for an amino acid sequence is not intact. The coding region is split into segments that code for protein, called **exons**, and segments that do not code for protein, called **introns**. The existence of introns was first discovered in 1977 when two research groups, sequencing different genes in different animals, both found that there were "gaps" in their sequences. These gaps were flanked by DNA that matched the expected sequences based upon a reverse translation of amino acid sequences using the genetic code (see below) but the DNA sequences in the gaps made no sense. It was soon found that many such gaps existed in genes sequenced in animals and plants. In fact, the only place where these gaps, now called introns, did not occur was in the genes of bacteria. A survey of animal and plant gene sequences also showed that all introns started with the DNA sequence GT and ended with the DNA sequence AG. It has also been found that, while most genes in plants and animals contain introns, not all genes do. These genes are simply referred to as **intronless** genes. In addition, the number and size of introns varies widely. In human genes, for example, some genes may only have one intron while some have been found that have dozens. The size range of introns is from as few as nine bases to more than one hundred thousand bases. The question that was raised with their discovery as to why introns exist and what their function may be is still not resolved.

Within the first and last exon of genes are sequences that are part of the coding region but do not code for protein. These are called the **5'-** and **3'- UTRs** for untranslated regions. The 5'-UTR, which can vary in size but is usually short, extends from the beginning of the first exon and ends at the **Start Codon**. The start codon is the DNA triplet ATG which is the only sequence that encodes the amino acid methionine. This is where all proteins begin. At the other end of the gene, the 3'-UTR begins with one of the three Stop codons (TAA, TAG, or TGA) and ends just past the DNA sequence known as the **Polyadenylation**, or **Poly-A**, **signal**. The poly-A signal in most genes is the six base

sequence AATAAA but there are other, minor poly-A signals. This sequence is used in producing the 3'end of a mature messenger RNA, or mRNA (see below).

Preceding the 5'-UTR is a region of DNA that is populated with a large family of DNA signals that are used to direct gene expression. Most eukaryotic genes are transcribed by the enzyme polymerase II (pol II). Activation of pol II requires the assembly of a large complex of **transcription factors** that are assembled in a precise order around a specific signal called the TATA-box (Figure 2). The TATA-box, or Goldberg-Hogness Box, is a sequence of DNA bases lying 25 bases in front of the point where transcription will start. Hawkins (1996) presents the consensus TATA-box sequence data for 680 eukaryotic genes. The name TATA comes from the fact that the sequence TATAAA is found in the vast majority of these consensus sequences and that the sequence TATATA is the second most common.

Transcription begins with a protein called TF-IID (transcription factor IID) binding to the TATA-box. TF-IID is a very large protein composed of a number of distinct subunits. Among these subunits is the part that binds to the DNA of the TATA-box while other subunits, called Transcription binding protein associated factors (TAFs), are required to mediate the activation of transcription. Once this complex is established, it is stabilized by TF-IIA. Next, TF-IIB is recruited and it interacts with both TBP and the DNA backbone. The amino terminal end of TF-IIB extends from the TATA-box to the transcription start site 25bases downstream and fixes the site. The rest of the TF-IIB molecule serves as a landing site for TF-IIF, pol II, and some additional TAFs. TF-IIF begins to unwind the DNA just ahead of pol II so that it can begin transcription. Once this starts TF-IIE, TF-IIH, and TF-IIJ bind as pol II starts to transcribe the gene. TF-IIH has the function of phosphorylating pol II after it is activated by TF-IIE. The phosphorylated form of pol II, called pol IIO, is then bound by TF-IIS and TF-IIX which regulate the rate of pol IIO transcription.

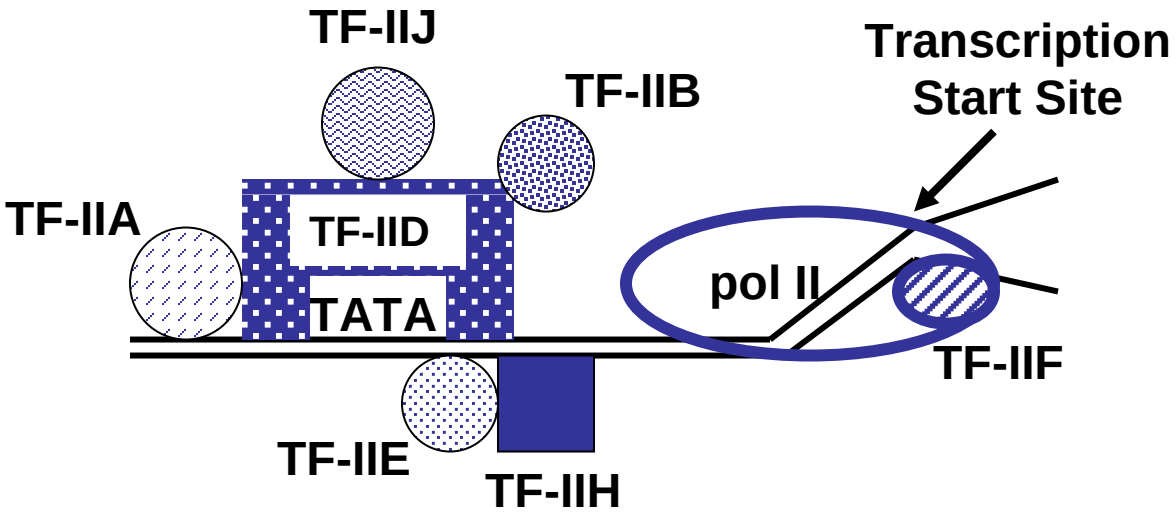


Figure 2. Schematic drawing of the transcription initiation complex.

In addition to these transcription factors there are many other factors further upstream that serve regulatory roles. Many of these also have specific DNA base sequences that they recognize and bind to. Two sequences, GCCAAT and GGGCGG, are the sites for proteins that enhance binding of TF-IIB. One of the proteins known to bind to the sequence GGGCGG is Sp1 and it can bind to either strand of the DNA. The group of proteins that bind to GCCAAT is collectively known as C/EBF or CBF (CAAT-box factors). Others with known binding sequences are AP1 (TGASTCA), CREB (TGACGTCA), CTF/NF-1 ($TGGCCN_5GCCAA$), and NF-κB (GGGRANTYYCC). In standard DNA nomenclature N means any of the four DNA bases, R means either G or A, Y means either T or C, and S means either G or C. An excellent web site that will identify all potential transcription and regulation factor sites in an input DNA sequence is found at http://bimas.dcrt.nih.gov/molbio/signal/.

The Primary Transcript

Once pol II is initiated in the presence of all necessary co-factors, promoters, and regulatory elements, it begins to make an **RNA copy** of the gene. This copy will include all of the sequence from the transcription initiation site through to the end of the 3'-UTR. This includes the introns as well as the exons. This all-RNA copy is called the **primary transcript**. During formation of the primary transcript three separate functions are initiated through which the primary transcript is **processed**. These three separate functions are **capping**, **intron excision**, and **polyadenylation**. Once all three are completed, the primary transcript has become a mature messenger RNA (mRNA) that is ready to leave the nucleus to carry out its function in protein translation of the genetic message (Figure 3).

The process of 5' capping begins as soon as primary transcription begins. An enzyme called guanyl transferase adds a guanyl radical to the 5' end of the nascent RNA transcript. This is, in turn, methylated by another enzyme called a methyl transferase. The cap, a 7-methyl-guanyl molecule, functions in binding mRNAs to the ribosomes. It also stabilizes the mRNA by protecting it from enzymes that might destroy it in the cytoplasm. Finally, the cap is the site of recognition of a protein called CBP (cap binding protein) that also aids in binding the mRNA to the ribosome.

The process of intron excision is another molecular ballet that rivals transcription in its intricacy and precision. Intron excision, shown in Figure 4, begins when two molecules, called U1 and U2, bind to the 5'-end and the 3'-end respectively of an intron. Binding is aided by a protein called PRP5 in the presence of the energy source adenosine triphosphate (ATP). U1 and U2 are molecular complexes composed of RNA and protein that belong to a class of molecules termed small nuclear riboproteins (snRNPs). U1 and U2 align themselves such that a loop is formed by the intron bringing the exons close together. At this time another snRNP complex, formed by U4, U5, and U6, binds to the adjacent U1 and U2 forming the structure called a **spliceosome**. Once more, in the

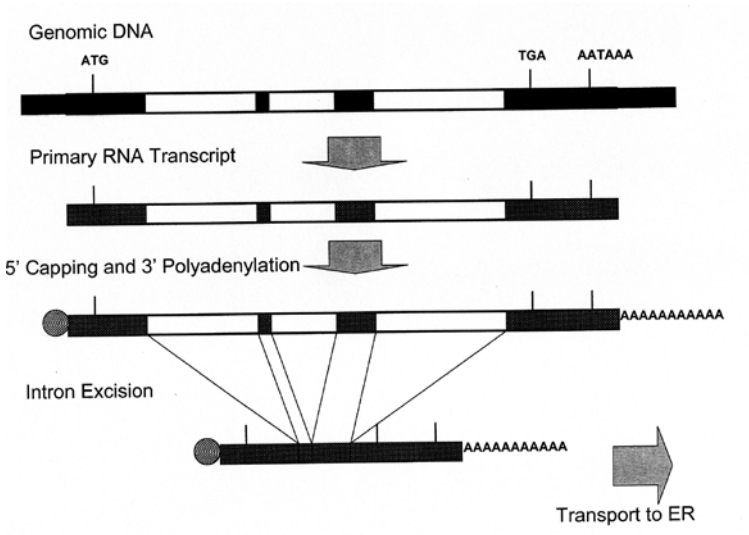


Figure 3. The process of transcription. The primary RNA transcript is generated via an RNA polymerase activated in concert with various regulatory and transcription factors. This primary transcript is then processed via 5' capping, 3' polyadenylation, and intron excision into a mature messenger RNA (mRNA). The mRNA is then transported to the endoplasmic reticulum where the message is translated on the ribosomes into protein.

presence of ATP as the energy source, a protein, PRP2, is recruited and the 5'-end of the intron is cleaved at the appropriate GU dinucleotide. The snRNPs U1, U4, and U6 then attach the 5'-end to a site within the intron forming a structure called a **lariat**. The lariat is cleaved off at the 3'-end AG dinucleotide in the presence of yet another protein, PRP16, and ATP. The free lariat intron, bound with all five snRNPs, is further processed into free ribonucleotides and the snRNPs are freed to process another intron. The exon ends are spliced together by PRP22 and the process is complete.

As with capping, intron excision begins as soon as there are introns available. The representation of intron splicing shown below in Figure 5 has been verified in photographs taken through an electron microscope and through experiments in which various components of the process are disabled. Disabling specific components by a process of site-directed mutation or site-directed mutagenesis in which a crucial part of the gene encoding the protein of interest is altered is generally termed a "knock-out." Knock-outs are a powerful means of assessing the function of gene products and have become a standard part of the molecular toolbox used by researchers.

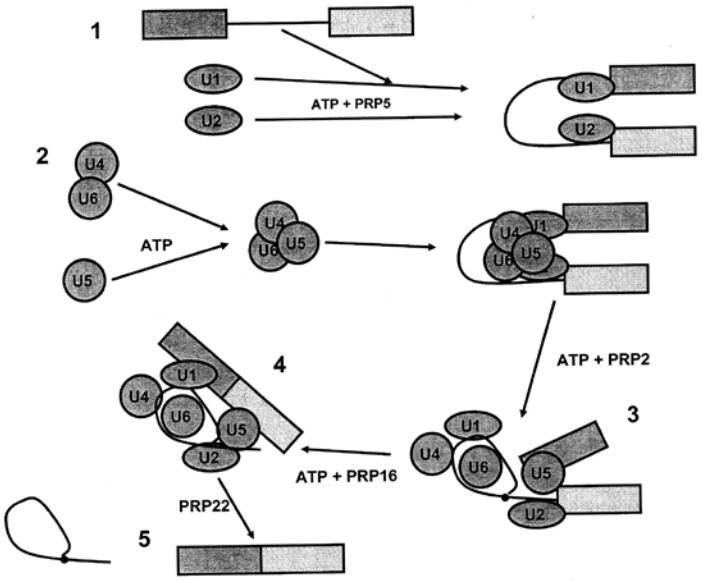


Figure 4. Schematic representation of the process of intron excision.

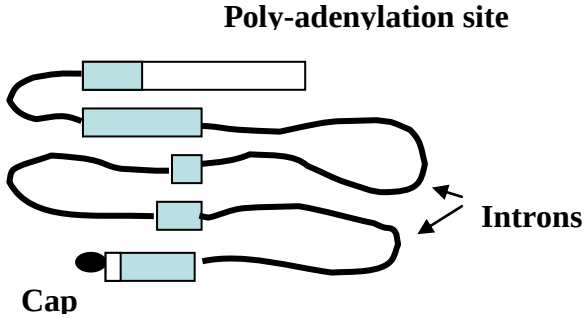


Figure 5. Schematic representation of mRNA processing showing alignment of 5'- and 3'-intron ends.

The last part of preparing a primary transcript for transport to the ribosomes is **polyadenylation**. In the 3' UTR of the primary transcript is a sequence AAUAAA. This is called the **polyadenylation signal**. Just 23 to 24 bases downstream is another sequence GU/U. These two sequences define the polyadenylation site. Polyadenylation begins when four proteins called CPSF (capping and polyadenylation specificity factor), CF-I and CF-II (cleavage factors 1 and 2), and CstF (cleavage stimulation factor) bind to the polyadenylation site (Figure 6). CF-I, CF-II, and CstF then cleave the RNA just downstream of the polyadenylation signal. CPSF remains bound to the RNA and an enzyme called PAP (poly-A polymerase) is recruited into a complex with the RNA and the CPSF. Once bound, PAP begins to extend the 3'-end of the RNA by adding only

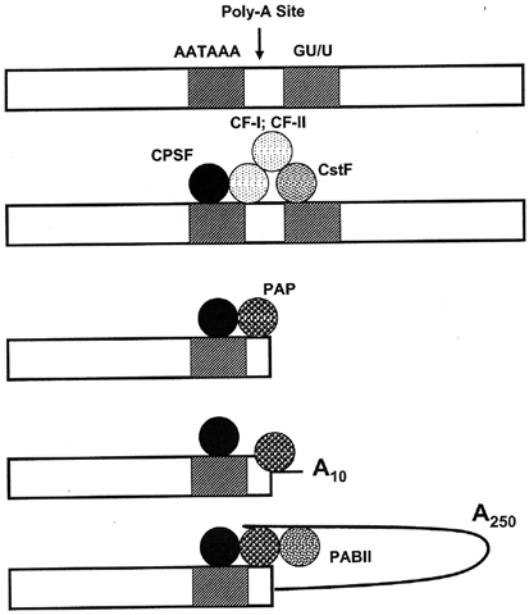


Figure 6. Schematic representation of the process of poly-adenylation.

adenosine nucleotides. This poly-A addition continues for at least 250 nucleotides during which another protein called PABII (poly-A binding protein) is recruited to form a loop back to the poly-A site. Once this process is complete, the mature mRNA is ready for transport from the nucleus to the ribosome where protein **translation** is carried out.

In a study of thousands of human mRNAs, Beaudoing et al. (2000) show that the classic AAUAAA poly-adenylation signal is present only about three-fifths of the time (58.2%). The next most common sequence is AUUAAA, present in 14.9% of the mRNAs observed. The remaining 26.9% of poly-adenylation sites are composed of a large number of sequences including AGUAAA, UAUAAA, CAUAAA, GAUAAA, AAUAUA, and so on. Beaudoing et al. also found that fully 23.3% of mRNAs have two or more polyadenylation signals. Further, when two or more poly-adenylation signals are present, the typical AAUAAA sequence is always the most common as the last sequence. That is, if there are three poly-A signals, the third is most often AAUAAA while the first two are mostly not AAUAAA. The same is true for two and four poly-A signals. Finally, the more poly-A signals there are at the 3'end of an mRNA, the longer the 3' UTR is. If there is only one poly-A sequence, the 3'UTR averages just over 500 nucleotides in length whereas 3' UTRs with four poly-A signals average just over 2000 nucleotides in length.

Translation

Translation of the genetic message into protein occurs in the ribosomes attached to the cellular structure called the **endoplasmic reticulum**. Ribosomes are themselves complex structures composed of RNA and protein. Ribosomes are synthesized in the nucleolus, a structurally and functionally specific substructure within the nucleus. Functional

ribosomes are composed of two subunits called the large, or 60S, subunit and the small, or 40S, subunit. The terms 60S and 40S refer to the sedimentation value in Svedberg units. The molecular weight of the 60S subunit is around 2.8 million Daltons while the molecular weight of the 40S subunit is around 1.4 million Daltons. Thus a complete ribosome has a molecular weight of 4.2 million Daltons. The components of the ribosome subunits are an 18S ribosomal RNA (rRNA) and 33 different proteins in the small subunit and three rRNAs, 5S, 5.8S, and 28S, plus 50 different proteins in the large subunit.

When the structure of the DNA molecule was published by Watson and Crick in 1953 (Nature, 171: 737-738, 1953), two questions were left to resolve. The first was the mechanism by which DNA replicated itself and the second, stated succinctly by Crick, was how a sequence of four things (the DNA bases) could encode a sequence of twenty things (the amino acids of protein). The search for the Genetic Code went on for more than a decade until, in 1966, a universal genetic code was announced in Volume 31 of the Cold Spring harbor Symposia on Quantitative Biology. Leading the search were H. Gobind Khorana, Severo Ochoa, Matthew Meselson, Marshall Nirenberg, and Heinrich Matthaei. The methods they used were tedious and labor intensive but success was inevitable. They determined that the genetic code was composed of three letter "words" called **codons** and each codon called for a specific amino acid. However, four DNA bases "read" three at a time gave 64 possible codons. Since there were only twenty amino acids, this meant that more than one codon could call for the same amino acid. This phenomenon is called **degeneracy**. Shown below in Figure 7 is the genetic code.

Figure 7. The genetic code. The bases shown are RNA bases since the code is actually read from the mature mRNA transcript.

Sixty-one codons actually call for an amino acid while three codons, UAA, UAG, and UGA are stop signals for protein translation. The degeneracy of the genetic code is more easily seen when presented by amino acid,

Ala	Arg	Asn	Asp	Cys	Glu	Gln	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
GCU	CGU	AAU	GAU	UGU	GAG	CAG	GGU	CAU	AUU	CUU	AAG	AUG	UUU	CCU	UCU	ACU	UGG	UAU	GUU
GCC	CGC	AAC	GAC	UGC	GAA	CAA	GGC	CAC	AUC	CUC	AAA		UUC	CCC	UCC	ACC		UAC	GUC
GCG	CGG						GGG		AUA	CUG				CCG	UCG	ACG			GUG
GCA	CAA						GGA			CUA				CCA	UCA	ACA			GUA
	AGG									UUG					AGU				
	AGA									UUA					AGC				

Here, it can easily be seen that three amino acids are called by six different codons, five amino acids are called by four different codons, one amino acid is called by three different codons, nine amino acids are called by two different codons, and two amino acids have but one codon each. In all cases except where there are six codons, the codons that call an amino acid will differ only in the third position. This was called the "wobble hypothesis" when first suggested by Francis Crick.

At this point the genetic code is known to be "read" at the ribosomes and the three letter codon "words" are used to select the appropriate amino acid. The final piece of this puzzle is how does the process of translation proceed. It is customary to divide the process of translation into three separate phases that occur sequentially. These phases are **initiation**, **elongation**, and **termination**.

The start, or initiation, codon is almost always AUG, which codes for the amino acid methionine (Met). As the mRNA approaches the ribosome, the 5' cap 7-methyl-guanyl molecule binds a large cap binding protein (CBP). This protein, in turn, binds to the 40S ribosomal subunit, setting the mRNA in place. At the same time a small RNA called transfer RNA (tRNA) binds the amino acid Met via a reaction catalyzed by an enzyme called amino acyl tRNA transferase. Each codon, with the exception of the three stop codons, has its own tRNA. The specificity of the tRNA for the codon is determined by a three base sequence that is complementary to the codon sequence. This is called the **anti-codon**. Transfer RNAs are small, between 75 and 90 nucleotides in length, and have a very precise structure (Figure 8). The 3'-end is slightly longer than the 5'-end and is composed of the sequence CCA.

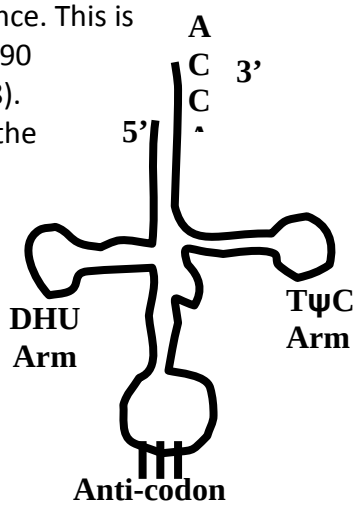


Figure 8. The structure of a tRNA molecule. The anti-codon loop lies at the opposite end from the CCA 3'-end that binds the amino acid. The other two arms are named for the non-standard RNA bases that are found there. The DHU Arm contains a dihydrouridine while the TψC Arm contains pseudouridine. The extra small arm shown between the anit-codon arm and the

TψC Arm is not present in all tRNAs.

Once the Methionyl-tRNA is present, the anti-codon aligns with the Met codon on the surface of the 40S ribosomal subunit. The 60S subunit then complexes with the 40S subunit and the mRNA (Figure 9). The initiation complex is now complete and the process of elongation begins.

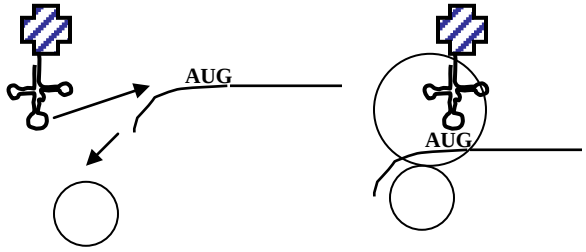
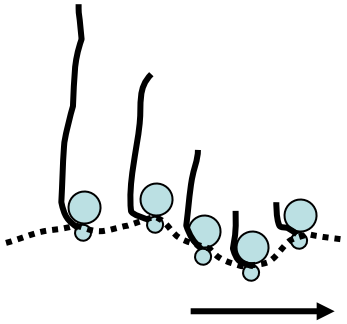


Figure 9. Formation of the initiation complex for mRNA translation and protein synthesis. On the left, the capped mRNA binds to the 40S ribosomal subunit and the Methionyl-tRNA anti-codon aligns with the AUG start codon. On the right, the 60S ribosomal subunit is in place and elongation begins.

Elongation is catalyzed by the rRNAs and a series of elongation factors that bring in the next tRNA in series and dissociate the previous tRNA. The process repeats until a stop codon is reached. At each step the tRNA will have its own amino acid as well as all of the previous amino acids in the protein bound through peptide linkages. As the mRNA passes through the ribosomes, the protein can be seen growing out of each ribosome (Figure 10).

Figure 10. Depiction of an mRNA (dashed line) threading through a series of ribsomes (a polysome) with amino acid chains (peptides) emerging from them. The arrow indicates the direction of movement of the mRNA through the polysome.



The final phase of the process of translation is termination. Termination begins when a stop codon enters the ribosome. There are no tRNAs that correspond to a stop codon and the finished peptide chain is liberated from the ribosome. There is a brief association of several soluble protein termination factors with the ribosome at this point. The last tRNA molecule is released and the two ribosomal subunits dissociate from the mRNA and each other whereupon they rejoin the pool of subunits to be used in other transcription and protein synthesis reactions.

Protein Structure

One final word about transcription and translation. The end product of DNA transcription and translation is an amino acid sequence, a poly-peptide. This is not, strictly speaking, a PROTEIN. Proteins are most often complex structures containing more than one poly-peptide chain or subunit. Even among those proteins that are composed of a single peptide chain, there is internal structure that is dictated by the specific order of the amino acids. Peptides all fold into some type of secondary and tertiary structures and it is these structures that confer function. There are binding sites, co-factors sites, and other structural regions within a protein that must be precisely configured for the protein to do its job. Much of this is carried within the genetic message but much of it is not. For example, a protein may need to fold into a specific shape for a co-factor to bind to it and the exact order and position of the amino acids required is carried in the genetic code for that protein. However, once the co-factor is bound to the protein it may undergo a further conformational change in its shape to attain its final functional form. This is not carried in the genetic code except indirectly as a consequence of the primary structure.

References

Beaudoing E, Freier S, et al. (2000) Patterns of variant polyadenylation signal usage in human genes. Genome Research, 10: 1001–1010.

Hawkins JD. (1996) Gene Structure and Expression. 3rd Ed. Cambridge: Cambridge University Press.

Resources

http://www.lewport.wnyric.org/jwanamaker/animations/Protein%20Synthesis%20-%20long.html

http://www.accessexcellence.org/RC/VL/GG/protein_synthesis.html

http://web.indstate.edu/thcme/mwking/protein-synthesis.html

http://www.brookscole.com/chemistry_d/templates/student_resources/shared_resources/animations/protein_synthesis/protein_synthesis.html

