As we've discussed, the structure of DNA provides a mechanism for self-replication. DNA also "stores" the genetic information that determines what a cell is and how it functions. In this section, we will look at how the information stored in DNA is used to direct the synthesis of proteins (specifically polypeptides) through the processes of transcription and translation. Ultimately, the specific proteins in a cell determine the expressed phenotype. In a later section we will look at some of the ways in which gene expression is regulated so that the appropriate DNA is expressed in each cell.

Before we discuss how DNA does its job of storing and using genetic information, let's look at some of the research that led to the conclusion that a gene is a piece of DNA that specifies the amino acid sequence in a polypeptide (or protein).

**Garrod's contribution**
In 1909 Archibald Garrod postulated that inherited diseases were caused by the inability of the individual to synthesize a particular enzyme as a result of a mutant allele. Garrod was the first to correlate "one gene to one enzyme". He was correct; many of our metabolic disorders are caused by not having a specific enzyme. However, it took decades of research to "prove" that a gene's function is to provide instructions on how to synthesize a specific protein, and that metabolic pathways for synthesis and degradation of molecules within a cell (and organism) are catalyzed by specific enzymes at each step of the pathway.

**Beadle, Ephrussi, Tatum, Drosophila eye color and Pink Bread Mold**
In the 1930's, George Beadle and Boris Ephrussi postulated that the variations in *Drosophila* eye color were caused by mutations in the enzymes that catalyzed eye pigment, but were not able to identify the chemical pathway of the enzymes.

A decade later Beadle and Edward Tatum induced mutations in *Neurospora*, a common pink mold, and tracked the nutrient metabolism of the mutant strains. *Neurospora* has a haploid life cycle so each individual has just one allele for each genetic trait.
They mapped chromosome locations of the mutant strains, and then related their chromosome maps to the presence or absence of specific enzymes needed in *Neurospora*'s metabolic pathway for the synthesis of *arginine*. They compared the wild type (prototroph) that could synthesize its nutrients from a minimal nutrient medium with several different auxotrophs (the mutant forms that could not survive on the minimal medium). From their research, Beadle and Tatum postulated the one gene-one enzyme theory.

Eventually, we also learned that not all genes must code for enzymes; some code for structural proteins or functional proteins (many of which we discussed in our first unit). Furthermore, quaternary proteins are composed of more than one polypeptide, so the concept of gene was further refined to be that a gene codes for a polypeptide.

Continued research found that some genes code for the many forms of RNA found in cells, and other genes are involved in gene regulation, controlling what DNA gets expressed in specific cells and tissues. At this time we shall discuss how the information stored in DNA is expressed in cells.
Information Flow from DNA to RNA to Protein
The expression of DNA occurs in two major processes:

- **Transcription**
  The information of a DNA nucleotide sequence is "copied", or transcribed, to its corresponding RNA nucleotide sequence, forming a molecule called messenger RNA.

- **Translation**
  The information of the RNA sequence is translated for the synthesis of the amino acid sequence of a specific polypeptide or protein at ribosomes.

Francis Crick worked on how genetic information gets from DNA to RNA (transcription) and what the relationship is between a specific nucleotide and a specific amino acid (translation) after he and James Watson determined the structure of DNA in the 1950s. Crick coined the phrase the central dogma to describe this molecular biology relationship of information flowing from DNA \(\rightarrow\) RNA \(\rightarrow\) Protein.

It should be noted that some viruses, which are non-cellular infectious particles, have RNA as their genetic molecule, hence are exceptions to the central dogma of living organisms.

Most RNA viruses have special virus encoded polymerases that use the viral RNA as a template for the host cell to make new virus RNA molecules as well as coding for the needed viral coating. Many plant viruses are RNA viruses, as are the viruses that cause colds and flu in humans.

Some RNA viruses are retroviruses that use reverse transcription. Their RNA is used as a template for the host cell to make a DNA strand that functions as a template to make a complementary DNA strand forming a double helix. This DNA is incorporated into the host cell DNA as a provirus or prophage (if the host is a bacterium). The provirus is transcribed and used to make more virus RNA, which can then form new viruses that leave the cell ready to infect more cells. HIV is a retrovirus. As an aside, DNA (genetic) technology takes advantage of the retrovirus reverse transcriptase enzymes to make recombinant DNA from RNA templates.
The Universal Genetic Code

The genetic code is common to life. The same DNA triplet codes for the same amino acid in almost all organisms from bacteria through plants and animals. We find few exceptions: Mitochondria and chloroplast DNA and some ciliate protists have unique amino acids and codons as do the Archaea. These codon differences may have been protective for endosymbionts early on in evolution.

The universality of DNA is one of the reasons for success in DNA technology work. A DNA sequence transferred from one organism to another can be transcribed, translated and expressed. Some of the earliest work in DNA technology tested fluorescent gene transfer. When successful, the gene’s expression was readily visible. Genes such as the fluorescent gene, whose expressions can be readily determined, are known as marker genes.

Although we will discuss, with some detail, the information flow from DNA to protein, first, let's look briefly at how the genetic code is read and interpreted and then learn about the different types of RNA used in transcription and translation.

Reading the Genetic Code

It's easy to say that the information needed to synthesize a polypeptide is stored in DNA. But how do DNA nucleotides store genetic information? And how does that information get from the nucleus where DNA is found to the cytoplasm where proteins are synthesized?

The Messenger Hypothesis and Transcription

Francis Crick and colleagues proposed that special RNA molecules were formed as a complement to one strand of the DNA molecule and used to carry a message: the messenger RNA (mRNA). Once transcribed, the mRNA moves to the cytoplasm as a nucleotide sequence of codons. As deciphered in the 1960's, each codon in the message (of RNA) is comprised of a three-nucleotide-long sequence (the triplet code). Each triplet sequence of nucleotides in a DNA molecule is a "code word" for one specific amino acid. The code is non-overlapping and lacks separators, or punctuation marks, between the triplets.

DNA molecules contain a linear sequence of nucleotide triplets that will specify which amino acids a protein will contain and the sequence, or order, in which these amino acids will peptide bond to form a polypeptide. The mRNA is a complement of the DNA sequence and delivers that transcript of instructions to ribosomes.
The "Adapter" Hypothesis and Translation
Crick also proposed that the process of translation, or the sequence of amino acid peptide bonding to synthesize a polypeptide (protein synthesis), would require an additional molecule that can bind to a specific amino acid and also recognize a specific triplet nucleotide sequence. That "adapter" molecule is tRNA.

Each tRNA has an amino acid binding site along its stem, which can attach to its specific amino acid. Specific enzymes do this. These attachment sites are also phosphorylated to provide the energy for protein synthesis.

Each tRNA molecule also has a triplet sequence that matches a triplet sequence of the mRNA. The only way to match nucleotides is by base pairs, which are complements to each other, so the tRNA triplet that codes for and attaches to a specific amino acid is called the anticodon. tRNA is the intermediate molecule that translates the sequence of nucleotide information of the DNA into a sequence of amino acids in a protein.

Codon-anticodon (mRNA-tRNA) matches occur at ribosomes where the amino acids, which are attached to the tRNA molecules, can be joined by peptide bonds to form polypeptides. Several ribosomes can function at one time so that several copies of a polypeptide can be made at one time.

Why a Triplet Code?
There are 20 different amino acids and just 4 DNA nucleotides. It was obvious that each nucleotide could not be a code word for a specific amino acid and taking the nucleotides in groups of 2 only results in 16 possibilities for amino acid code words. Taking the four nucleotides in groups of 3 results in 64 possible triplets, more than needed for the 20 amino acids, so researchers started their search for the code with the hypothesis that it would be a triplet code. They were correct.

To decipher the first code word, Marshall Nirenberg and JH Matthaei synthesized artificial RNA molecules comprised of just one nucleotide. When they used a synthetic poly-U in a mixture of ribosomes, amino acids and protein-synthesizing enzymes, a polypeptide consisting of just phenylalanine was formed. Poly-A resulted in a polypeptide of just lysine. Eventually codes for all of the amino acids were determined in a similar fashion to their work, although the code words that had mixed nucleotides required more elaborate procedures.

Although there can be 64 different DNA code words, three of them do not code for specific amino acids. These three code words specify the end of a polypeptide coding and are known as the "stop" code words or termination signal. A fourth code word, the triplet TAC, which codes for the amino acid methionine, is the "start" code that initiates translation at the ribosomes.
A mRNA nucleotide triplet (synthesized from the DNA template) that codes for a specific amino acid is the **codon**. The codon is a complementary (rather than identical) triplet to the DNA code word. Synthesis of RNA follows the same nitrogen base pair rules that dictate DNA duplication, with the exception that in RNA, uracil nucleotides pair with adenine. RNA does not have thymine. RNA molecules are also synthesized in the 5′ → 3′ direction reading a 3′ → 5′ DNA template.

The impact of deleting one, two or three nucleotides on gene transcription was also studied when scientists were "breaking" the DNA code. With one or two deletions, the portion of the gene transcribed past the point of the deletion was nonsense. This phenomenon is known as a **frame-shift** alteration of the genetic code. *(Frame-shift mutations will be discussed later.)* With three deletions, gene transcription was restored, but the protein might still be non-functional. This research confirmed that validity of the triplet code.

The DNA coding sequence for a single protein will have 3 times the number of nucleotides as the number of amino acids in the protein for which the sequence codes. In addition, there will be start and stop regions of the DNA associated with the instructions for the protein to be synthesized, and regions within the DNA molecule that do not code and are removed by RNA processing after transcription.

The triplet code is most often reproduced in Codon tables. When you look at a codon table, you see that some amino acids are coded for by more than one codon. Often, only the first two nucleotides of the triplet are essential; the third is redundant. *(e.g., CCU, CCC, CCA and CCG all code for the amino acid, proline, and UCU, UCC, UCA and UCG all code for the amino acid, serine.)* But the codon always contains the entire triplet. This appears to be an artifact of having more triplet possibilities than needed to code for 20 amino acids.

The reverse is not true. One codon never codes for more than one specific amino acid. UCU codes for serine. UCU does not code for any other amino acid.
RNA – The molecule that uses the information stored in DNA
Prior to discussing transcription and translation, it's valuable to look at RNA, our second nucleic acid, which has significant roles in gene expression.

Structure of RNA Molecules
RNA is composed of
- Phosphate
- Ribose sugar
- Four nucleotides
  - Adenine
  - Guanine
  - Cytosine
  - Uracil (Uracil replaces the thymine found in DNA)

Molecules of RNA are single-stranded. However, some RNA molecules fold back on themselves at places forming complementary base pair Hydrogen bonds. The double stranded sections, called hairpins, provide stability to the structure of the RNA molecule for its function in protein synthesis and provide resistance to the RNA hydrolyzing enzymes in the cytosol. In bacteria, a hairpin turn at the end of a RNA molecule being transcribed stops transcription.

There are a variety of types of RNA molecules. Although we are most familiar with those involved in transcription and translation, RNA molecules are also involved with regulating gene expression, and several of these RNA molecules will be discussed in our section on gene regulation.

<table>
<thead>
<tr>
<th>Type of RNA</th>
<th>Functions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Messenger RNA (mRNA)</td>
<td>Carries information specifying amino acid sequences of proteins from DNA to ribosomes.</td>
</tr>
<tr>
<td>Transfer RNA (tRNA)</td>
<td>Serves as adapter molecule in protein synthesis; translates mRNA codons into amino acids.</td>
</tr>
<tr>
<td>Ribosomal RNA (rRNA)</td>
<td>Plays catalytic (ribozyme) roles and structural roles in ribosomes.</td>
</tr>
<tr>
<td>Primary transcript</td>
<td>Serves as a precursor to mRNA, rRNA, or tRNA, before being processed by splicing or cleavage. Some intron RNA acts as a ribozyme, catalyzing its own splicing.</td>
</tr>
<tr>
<td>Small nuclear RNA (snRNA)</td>
<td>Plays structural and catalytic roles in spliceosomes, the complexes of protein and RNA that splice pre-mRNA.</td>
</tr>
<tr>
<td>SRP RNA</td>
<td>Is a component of the signal-recognition particle (SRP), the protein-RNA complex that recognizes the signal peptides of polypeptides targeted to the ER.</td>
</tr>
<tr>
<td>Small nucleolar RNA ( snoRNA)</td>
<td>Aids in processing of pre-rRNA transcripts for ribosome subunit formation in the nucleus.</td>
</tr>
<tr>
<td>Small interfering RNA (siRNA) and microRNA (miRNA)</td>
<td>Are involved in regulation of gene expression.</td>
</tr>
</tbody>
</table>
Messenger RNA (mRNA)

- mRNA is the unique blueprint, or transcript of a DNA gene sequence for each protein to be assembled in a cell. Each specific mRNA nucleotide sequence of three nucleotides carrying one bit of information for one amino acid in the protein to be synthesized. These mRNA triplets are the **codons** described by Crick’s colleagues (see later).
- mRNA is manufactured by transcription on demand from a template DNA strand in the nucleus that has instructions for the specific protein needed in the cell. There are a host of transcription factors that locate the appropriate DNA and initiate transcription.
- A specific mRNA molecule migrates from the nucleus to ribosomes for the process of translation.

Ribosomal RNA (rRNA)

- rRNA is a structural component, with protein, of the **ribosomes**.
- A ribosome is composed of 2 subunits, a small subunit containing RNA molecules plus stabilizing proteins, and a larger subunit containing RNA plus proteins and the enzymes/ribozymes needed for protein synthesis.
- The ribosomal subunits are manufactured in the nucleolus, but the complete ribosome, which forms the **ribozyme** complex, is found in the cytoplasm, frequently attaching to rough endoplasmic reticulum.
- The small rRNA subunit has a binding site for mRNA molecules during protein synthesis and a binding site to the large subunit. The larger subunit has three attachment sites for tRNA molecules, the P site (Peptidyl-tRNA site), the A site (Aminoacyl-tRNA site) and E site (Exit site), plus a transfer site where the tRNA first connects to the ribosome.
- The rRNA of the ribosome serves as a catalyst for peptide bond formation, hence the ribozyme label. The proteins serve to maintain the conformation of the ribosome.
Transfer RNA (tRNA)
- There are a number of different specific tRNA molecules in the cytoplasm.
- Each tRNA is about 80 nucleotides long and has 3 hairpin loops in which the RNA is folded back on itself making hydrogen bonds between complementary nucleotides. A fourth "leg" also base-pairs where the opposite ends of the tRNA molecule are adjacent to each other.
- Each different tRNA has two important pieces:
  - An amino acid attachment site at the 3' end, which can attach to one specific amino acid
  - A tRNA triplet sequence which pairs with one specific nucleotide triplet sequence found on the 2nd hairpin loop. This tRNA triplet specifies the precise amino acid that will attach to the attachment site of the tRNA.
- The tRNA triplet that determines the amino acid attachment is called the **anticodon**. It is the complement to a specific mRNA codon triplet.
- tRNA is the connection or interpreter between the information carried on the DNA gene and the amino acids that will be assembled into proteins. tRNA molecules carry their specific amino acids to the ribosome for peptide bonding to form the protein specified by the mRNA instructions.

Other kinds of RNA (to be discussed during the details of transcription and translation or in our section on gene regulation):
- **pre-RNA** is the primary RNA transcript in eukaryotes transcribed from DNA prior to being processed into a functional mRNA molecule.
- **SRP RNA**, (Signal Recognition Particle RNA) is part of the signal-recognition particle that moves ribosomes from the cytosol to the ER during translation.
- **snRNA** (small nuclear RNA) is formed from hairpin loops of 21-28 nucleotides. snRNA functions in gene regulation and in mRNA processing.
  - snRNA is a component of **snRNP** ("snurps"), small nuclear ribonucleoproteins that identify the 5' end of introns, mRNA removed in mRNA processing. snRNPs, with additional proteins, form **spliceosomes**.
  - Some snRNA binds to DNA altering DNA ability to be expressed.
- **siRNA** (RNAi), or interference RNA, combines with complementary sequences on mRNA inhibiting gene expression.
- **miRNA**, or microRNA, binds to mRNA transcripts that have complementary nucleotide sequences, blocking translation, thereby silencing the gene.
- **piRNA**, or piwi-associated RNA appears to help reestablish appropriate methylation patterns during gamete formation.
The expression of DNA in the genetic control of the cell – Recap

- DNA stores the instructions for each cell to function, precisely coded in its four-nucleotide "alphabet", A, T, C, and G. DNA also has regions that function in gene regulation, regions of no apparent function and regions that appear to be genetic "gibberish".
- DNA instructions direct the synthesis of polypeptides. Many become the enzymes that catalyze metabolic activities of cells. Others are structural proteins.
- DNA also contains instructions for the synthesis of the RNA molecules that function in protein synthesis and in gene regulation.
- DNA is not used directly as a template for protein synthesis, a process that occurs at ribosomes in the cytoplasm. DNA never leaves the nucleus.
- DNA is used as a template to build a set of RNA instructions, a process called transcription. The mRNA (messenger RNA), carries the genetic instructions from the nucleus to ribosomes in the cytoplasm.
- Translation occurs at ribosomes. During translation, the information carried by mRNA molecules is used to direct the assembly of specific amino acids into proteins. Transfer RNA and ribosomal RNA are needed for translation.
Transcription – the Details
RNA synthesis uses DNA as a template, and occurs in the nucleus. There are three stages in transcription: **Initiation**, **Elongation** and **Termination**. For a given gene, the 5'→3' DNA strand that will be read by RNA polymerase is called the **template strand**. The opposite strand, which is not read, is called the **non-template, or coding, strand**, because its sequence will be the same as the mRNA transcript being formed, with the exception of thymine on the DNA coding strand, and Uracil on the RNA transcript. Transcription is used to make all forms of RNA, although we shall focus on the transcription of mRNA in our discussion.

**RNA Polymerase and Transcription**
The enzyme that catalyzes RNA synthesis from the DNA template strand is **RNA polymerase**. In eukaryotes, there are at least three forms of RNA polymerase, one for each kind of RNA molecule. RNA polymerase II and III are also used for some of the small RNA molecules. Prokaryotes have just one RNA polymerase. RNA polymerase does not need a primer and has no proofreading function.

RNA polymerase II transcribes mRNA. RNA polymerase is one of the most complex enzymes known, with binding sites for:
- regulatory proteins
- the DNA template strand
- the RNA nucleotide subunits
- the promoter region of the gene to be transcribed

**Transcription Initiation**
- The region of DNA that codes for the specific gene to be transcribed, called the **transcription unit**, starts to unwind using the enzyme, RNA polymerase, to initiate the transcription process.
- Transcription is started at a region of the DNA molecule called the **promoter**, a specific DNA base sequence at the 3' end of each gene. A promoter determines the template strand of the DNA, where transcription will start and the direction of transcription.
Special proteins, called transcription factors, help RNA polymerase find the promoter regions on the DNA. Many signal transduction pathways, important in cell communication and cell signaling, result in the production of transcription factors. (How transcription factors function is discussed in our gene regulation section.)

The promoter "DNA/RNA polymerase/transcription factors" complex is called the transcription initiation complex. One core promoter for RNA polymerase II includes a specific DNA sequence called the TATA box, comprised of the nucleotide sequence TATAAA. Transcription factors recognize and bind to the TATA box of the core promoter. A second core promoter region is the CAAT box, found in most genes. General transcription factors bind to the CAAT box to assemble the initiation complex. (The initiation complex is an important research area for gene regulation, and will be revisited in that section.)

Rate of transcription is related to the efficiency of promoters. Some bacterial promoters facilitate very rapid transcription (every 2 seconds) while weaker promoters may transcribe every 10 minutes. Weaker promoters may have alterations in their recognition sequences in the TATA box region.
**Elongation**

RNA polymerase will move in the 3' to 5' direction along the DNA template during what is now called the *elongation* process of transcription. Like DNA, RNA is synthesized in the 5' to 3' direction from the 3' to 5' DNA template. About 10 – 20 DNA nucleotides are opened at a time for RNA polymerase to work.

RNA Nucleotides are added to the chain according to the complementary base pairing; that is:

- RNA A – DNA T
- RNA U – DNA A
- RNA C – DNA G
- RNA G – DNA C

Several copies of RNA polymerase can be present so that several RNA transcripts can be made of the gene (DNA sequence) at one time. As one mRNA is being transcribed, a new RNA polymerase molecule attaches to its transcription factors at the promoter and starts transcribing a second. As the second starts elongating, a third RNA polymerase can attach, until many, many mRNA molecules are being synthesized along the DNA template. This is especially common in bacteria.
**Termination**

There is a **terminator** sequence that tells the RNA polymerase to stop. The actual mRNA termination and separation of the RNA polymerase molecule from the DNA template may be several nucleotides beyond the nucleotide stop sequence in eukaryote organisms.

A particular signal sequence of AAUAAA, called the polyadenylation sequence, precedes the terminator sequence, which often forms a hairpin loop in the mRNA followed by a poly-U sequence, a weak pairing partner for the DNA template. The hairpin forces RNA polymerase to pause, and the weak U-A bonds dissociate freeing the pre-mRNA molecule. RNA polymerase is freed from the DNA molecule several nucleotides after the mRNA, often by enzymes or helper proteins that are digesting the newly post-termination RNA molecules.
Modifying the pre-RNA for use: RNA Processing

In eukaryotic organisms, the mRNA initially transcribed is called the pre-mRNA transcript and requires modification before leaving the nucleus.

Cap and Tail
- A cap, called the 5' or G cap, made of methylated GTP, is attached to the 5' end of the RNA molecule (the end made first). A short untranslated mRNA region (UTR) follows the cap.
  - The cap protects the mRNA from being degraded by ribonucleases and helps in translation by attaching the mRNA transcript to the ribosome.
- At the end of transcription, a 3' poly-A tail of 50 - 250 adenine nucleotides is attached at the polyadenylation nucleotide sequence (AAUAAA). The polyadenylation sequence signals an enzyme to cut the pre-mRNA. After the cut, a second enzyme adds the adenine nucleotides. The tail
  - The tail facilitates the movement of the mRNA transcript through the nuclear pores and stabilizes the mRNA.

RNA Splicing – Introns and Exons

The primary transcribed eukaryote RNA molecule, called pre-RNA, consists of far more nucleotides than are actually used in protein synthesis. Some parts of the transcribed gene, called introns, do not code for amino acids in the protein to be synthesized. The name intron is derived from the fact that the introns are intervening segments that interrupt the message.

The discovery of introns used a technique called DNA or nucleic acid hybridization, one of the techniques common in genetic technology research. Although working independently, Richard Roberts and Phillip Sharp won the Nobel Prize in 1993 for this discovery. (Nucleic acid hybridization will be discussed in a later section.)

The regions that code are called exons because they are expressed. (The cap and tail are not considered introns or exons, but are a critical part of the final mRNA transcript.) As much as 90% of the pre-mRNA transcript may be non-protein coding nucleotides. Only about 25% of the entire genome is even coded at all. In humans, exons may comprise as few as 1200 of the average 27,000 nucleotides of a pre-mRNA (1% of human DNA).

Prior to use, introns must be removed from the pre-mRNA transcript, which is done during the RNA processing stage. This process is called RNA splicing, because the introns are cut out and the remaining exons get spliced together.
The process of gene splicing takes place at spliceosomes in the nucleus. Small pieces of RNA-protein complexes, called snRNPs or "snurps" (for small nuclear ribonucleoprotein particles) recognize specific codes, called consensus sequences, on the pre-mRNA at both ends of the introns. Many genes have such sequences at their intron-exon boundaries. The genetic disease, \( \beta \)-thalassemia, is caused by a mutation in a consensus sequence so the polypeptide is not spliced correctly.

Several snRNPs aggregate with additional protein to form the spliceosome bodies in an endergonic reaction.

The small nuclear RNA molecule portion (snRNA) (comprised of about 150 nucleotides) of the snRNPS serves as the catalyst for the intron excision and splicing of the exons, as well as in the formation of spliceosomes and the exon splice sites. Although our familiar catalysts in living organisms are enzymes, some RNA molecules, including snRNPs, function as catalysts. Such RNA molecules are called ribozymes. The base-pairing capability of the single stranded RNA facilitates its role as a catalyst.

Ribozymes also process pre-RNA for ribosomal RNA (rRNA) in some organisms. In these cases, the intron itself is the ribozyme, and the introns are removed by "self-splicing". Introns found in prokaryotes are also self-splicing. (Introns are rare or absent in prokaryote DNA.)

The completed mRNA (with a precise linear sequence of nucleotides) can now move from the nucleus to the cytoplasm for translation.

With the number of modifications that can occur to a mRNA transcript, one "gene" can, and some cases, does code for more than one polypeptide. This may be why our 30,000 human genes may code for the thousands of different proteins.
Why Introns?
An area of interest in science is why so much of the DNA molecule contains introns. There is much research in this field.

• We know for some genes, one pre-mRNA can be used to code for more than one protein, depending on what is determined to be introns in the processing stage. This is called **alternative gene splicing** and is fairly common. The proteins that determine gender development in fruit flies have been shown to share a common pre-mRNA. Alternative gene splicing is also used in antibody formation and in the synthesis of transport proteins.

- Introns may help in modifying protein shape. Different introns can affect the location of an active site for an enzyme or the attachment site for a membrane protein. This may affect change in proteins through time and may be involved with protein **domains**. For example, the active site is one protein domain. The attachment site for a co-factor is a second domain, and the attachment site for an enzyme to a membrane surface is a third domain. Introns can be used to modify domains without having the remainder of the protein's instructions recoded.

- Some introns may regulate transcription. There are some genes that are inhibited when RNA binds to the DNA molecule.

- Introns may help regulate the movement of the RNA from the nucleus to the cytoplasm of the cell.

- It is also believed that length of introns affects rate of recombination. Long introns give more room for crossing over of intact exons, hence more genetic variation. This is called **exon shuffling**.
Protein Synthesis – The Process of Translation

Once we have a final mRNA transcript, the mRNA is moved from the nucleus of the cell to the cytoplasm where it attaches to the small subunit of a ribosome in preparation for the process of translation. Translation also involves Initiation, Elongation and Termination steps.

Translation is where the information coded in DNA molecules is interpreted and translated to direct the actual synthesis of proteins, the function of tRNA molecules. To be successful, a tRNA molecule must be able to read and match the mRNA codon (that has the DNA nucleotide message) and carry its specific amino acid that corresponds to the appropriate mRNA codon to the ribosome. tRNA anticodons temporarily bond to their matching codon at the ribosome to facilitate the peptide bonding of the amino acids in the order determined by the original DNA nucleotide sequence.

Amino Acid Attachment

Prior to translation, one additional activity must occur: amino acids must be attached to their appropriate tRNA molecules. Our cells maintain a pool of the twenty amino acids and the permanent tRNA molecules at all times. The process of amino acid attachment involves ATP and a set of aminoacyl-tRNA synthetase enzymes. ATP loses two of its phosphates in the process and AMP is complexed to the amino acid-tRNA, forming a charged tRNA. Both the shape and charge of the tRNA molecules and the amino acids are important for the correct recognition and attachment of its specific aminoacyl-tRNA synthetase enzyme.

The amino acid activating enzymes must correctly interpret the genetic code and attach the appropriate amino acid to its corresponding tRNA. There are 20 different activating enzymes, one for each of the different amino acids.

Although theoretically there should be 61 different tRNAs, one for each triplet code word except for the 3 stop triplets, there are only about 45. As mentioned, the DNA triplet code for amino acids is redundant. Often the third nucleotide is not crucial so there is no need for 64 tRNA molecules.
Some tRNA molecules have a modified adenine as the third base, called inosine, that can base pair with any of the other nucleotides. These tRNA molecules recognize more than one mRNA codon. The tRNA molecule that attaches to the amino acid, leucine, for example, recognizes a codon that is CU_. The third nucleotide on the mRNA codon does not matter. When the third tRNA base is uracil, the tRNA can bond to either adenine or guanine as the mRNA codon's third base. This tRNA "flexibility" is called wobble.

**Translation Details - Initiation**

Initiation brings mRNA, the first tRNA and the two ribosomal subunits together.

- The small rRNA subunit has a binding site for mRNA molecules during protein synthesis and a binding site for the initial tRNA. The larger rRNA subunit has three attachment sites for tRNA molecules, the P site (Peptidyl-tRNA site), the A site (Aminoacyl-tRNA site) and E site (Exit site), plus the transfer site where the tRNA is brought to the ribosome by its protein escort. During protein synthesis the two subunits bind together.

- To initiate protein synthesis in eukaryotes, protein initiation factors bind the 5' leader of the mRNA to the small ribosome. The small ribosomal subunit migrates along the mRNA until it reaches the start codon, which is always AUG. The tRNA that has the "initiator" anticodon, UAC and its amino acid, methionine, attaches to the start codon of the mRNA that is located some distance beyond the 5' cap.

- The large ribosomal subunit binds to the small subunit, and the initiator tRNA attaches to the P site of the large subunit of the ribosome with the assistance of protein initiation factors, bringing the complex together and forming a functional ribosome. GTP provides the energy for the initiation process.

- A functioning ribosome is large enough to hold three mRNA codons. As stated, the first tRNA with its amino acid attaches to the P site. The A site of the larger subunit will be available for the 2nd tRNA molecule's anticodon to bind to the 2nd mRNA codon during elongation. The third codon site is the exit site.

- Note: Polypeptide synthesis is initiated at the amino end of the chain. Amino acids can only be added to the carboxyl end of an amino acid on the ribosome. Ribosomes move along the mRNA from 5'→3'
Translation Details – Elongation
Elongation involves three activities: **codon recognition** by tRNA molecules, **peptide bonding**, and **translocation** of the ribosome.

- **Codon Recognition**
  The 2nd tRNA molecule, with its attached amino acid, is brought into place at the ribosome's **A site** with the assistance of proteins called **elongation factors** and 2 molecules of **GTP**, which provide energy, as determined by the mRNA codon message. The tRNA anticodon hydrogen bonds to the mRNA codon at this time.

- **Peptide Bonding**
  The positioning of the two tRNA molecules (each with its proper amino acid) at the P and A sites is such that a peptide bond can be formed between the two amino acids that are attached to their respective tRNAs. rRNA functions as a ribozyme to catalyze the peptide bond between the P site's amino acid's carboxyl and the A site's amino acid's amino functional group at the peptide bonding A site on the ribosome. The catalyst ability of the large subunit's rRNA is called **peptidyl transferase**.

  This process also detaches the P site amino acid from its tRNA; the A site amino acid has its carboxyl group still attached to its tRNA along with the dipeptide. The polypeptide chain always elongates at the A site. The initial amino acid is known as the N-terminus of the polypeptide, and the final amino acid is the C-terminus.

- **Translocation**
  Once the peptide bond is formed, the ribosome will shift the A-site tRNA to the P site and the P site tRNA will shift to the E site to be dislodged from the ribosome (which is why the E site is called the exit site). **GTP** is also required for the translocation process. Because the mRNA is attached to the tRNA on P site, the mRNA is moved along the ribosome with the tRNA molecules to locate the 3rd mRNA codon at the now vacant A site. The 3rd tRNA will be brought into the A site by elongation factor proteins. The first tRNA can now be activated with another methionine in the cytoplasm.

  The codon-anticodon binding, peptide bonding, detachment of tRNA and shifting continues until all of the codons of the mRNA have been matched by tRNA anticodons. Note that the mRNA moves along the ribosome with its 5' end leading. mRNA moves only in one direction. Ribosomes and mRNA move relative to each other, codon by codon, unidirectionally.
Translation – Elongation of the Polypeptide
Translation – Termination

- When the mRNA stop codon (UAA, UAG or UGA) reaches the A site, additional tRNA molecules are prevented from attaching to the A site. A releasing factor protein attaches instead and hydrolyzes (adds a water molecule) the polypeptide causing it to be released from the ribosome tunnel.
- The ribosomal subunits and related proteins dissociate with the assistance of addition releasing factors using more GTP.

<table>
<thead>
<tr>
<th>TRANSCRIPTION</th>
<th>TRANSLATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initiation</td>
<td>Promoter DNA</td>
</tr>
<tr>
<td>Termination</td>
<td>Terminator DNA</td>
</tr>
</tbody>
</table>

Signal Codons for Transcription and Translation
Translation Summary

Amino Acid Attachment

Initiation

Leader sequence

Initiation factor

Large ribosomal subunit

Small ribosomal subunit (containing ribosomal RNA)

Initiation complex

Elongation

Elongation factor

Polypeptide chain released

Termination

Val Ser Release factor

Val Ser Ala Trp

Polypeptide chain released

Large ribosomal subunit

Small ribosomal subunit
Summary of Transcription and Translation

1. In the cell nucleus, RNA polymerase transcribes RNA from DNA.

2. Primary RNA transcript. Introns are excised from the RNA transcript, and the remaining exons are spliced together, producing mRNA.

3. mRNA is transported out of the nucleus. In the cytoplasm, ribosomal subunits bind to the mRNA.

4. tRNA molecules become attached to specific amino acids with the help of activating enzymes. Amino acids are brought to the ribosome in the order directed by the mRNA.

5. tRNAs bring their amino acids into the A site on the ribosome. Peptide bonds form between amino acids at the P site, and tRNAs exit the ribosome from the E site.

6. The polypeptide chain grows until the protein is completed.
Regulating Translation
Enhancing the Rate of Translation
A polypeptide is generally synthesized in about a minute. However, it is typical of mRNA to be working along many ribosomes at a time to direct the synthesis of many polypeptide molecules in sequence. As soon as the 5' end of a mRNA leaves one ribosome it will attach to the small subunit of an adjacent ribosome to initiate protein synthesis at that ribosome so one mRNA will be attached to many ribosomes at once. Such complexes are called polyribosomes or polysomes.

Inhibiting Translation in Pathogenic Prokaryotes
Both prokaryotes and eukaryotes have ribosomes, but the structure of ribosomes is slightly different in the two kinds of organisms. This difference is used in the development of antibiotics that target ribosomes, which blocks translation.
A number of antibiotics target some stage of protein synthesis in the prokaryote.

Recall that other successful antibiotics target cell wall synthesis, inhibiting peptidoglycan production, a substance found only in bacterial cell walls.

Effect of Penicillin on Cell Wall Synthesis
Targeting Proteins to their Functional Location

mRNA initially attaches to free ribosomes (and polysomes) in the cytoplasm. Proteins that will function in the cytosol are synthesized at free ribosomes, and when completed remain in the cytoplasm.

But many proteins will function in organelles or will be transferred to the ER for completion and modification. A protein destined for mitochondria, chloroplasts, nuclei and other organelles after translation will have an oligopeptide signal sequence within the polypeptide that is recognized by a docking protein in the membrane of the target organelle.

Polypeptides that are associated with the endomembrane system or destined for export from the cell of have a signal peptide sequence (of about 10 - 30 nucleotides) located near the start of the growing polypeptide that is recognized by signal-recognition particles (SRPs) in the cytoplasm. SRPs are RNA-protein particles. SRPs bind to and move the polypeptides to the rough ER for completion early during translation. The SRP brings the ribosome to a receptor protein site on the ER. As protein synthesis continues on the attached ribosome, elongating polypeptides move through ER pores into the ER cisternal spaces.
Modifying Polypeptides into Functional Proteins
As discussed in earlier units, the secondary, tertiary and quaternary structure of proteins follows the polypeptide synthesis to obtain a functional protein conformation that may involve [chaparons](https://www.example.com) in the cytoplasm. Other post-translation modifications are also common.

Post Translation Protein Modification
Most proteins undergo some form of modification after polypeptide synthesis prior to their functional conformation. Much of this happens in the endoplasmic reticulum in response to signal sequences in the completed polypeptide.

- Carbohydrate portions may be added to the protein forming glycoproteins, a process called [glycosylation](https://www.example.com). Glycosylation occurs in the ER or in the Golgi complex.
- Carbohydrate portions may direct proteins to their destinations, such as migration of hydrolytic enzymes to lysosomes.
- Carbohydrate portions stabilize storage proteins, especially in seeds.
- Membrane glycoproteins are important in cell recognition and membrane structure.
- Proteins may undergo [phosphorylation](https://www.example.com), catalyzed by protein kinases. (A kinase is, by definition, an enzyme that phosphorylates.) Phosphorylated proteins change shape which alters function by exposing a binding site or perhaps making the active site of an enzyme available.
- Proteins may also be subjected to [proteolysis](https://www.example.com), a process that fragments the polypeptides or removes non-functional peptide sequences.
  - Signal peptide sequences used to carry proteins to the ER need to be removed.
  - Methionine, the first amino acid of virtually all proteins, is usually removed.
  - Some proteins are synthesized in inactive forms deliberately, such as most hydrolytic enzymes. They will be converted to their active form when they reach their target location.
  - Proteases cleave polyproteins (long polypeptides that are synthesized as a unit, but are actually a set of proteins) into their individual functional proteins. Viruses contain proteases. One HIV drug target is the HIV protease.
Comparing Gene Expression in Prokaryotes and Eukaryotes

Although we have been discussing protein synthesis as a general concept, there are some differences between protein synthesis in eukaryotes and in prokaryotes.

- The prokaryotic RNA polymerase is different in structure from the eukaryotic RNA polymerase and does not depend on transcription factors.
- In prokaryotes, the two ribosomal subunits have a special bonding site, called the Shine-Delgarno site, after which the start codon aligns with the P site of the large subunit. Eukaryotic mRNA first binds to the small subunit of its ribosome.
- Prokaryote translation begins with N-formyl-methionine (fMet) instead of methionine in eukaryotes.
- Prokaryotes do not add a 5' GTP cap as is found in the eukaryotes.
- As stated earlier, introns, if present in prokaryotes, are removed during transcription so virtually no mRNA processing is required prior to translation.
- Prokaryote mRNA molecules may carry code for many proteins, not just one polypeptide as found in eukaryotes.
- Because prokaryotes have no separation of nucleus and cytosol, translation can be initiated as soon as the growing mRNA transcript is freed from the DNA.