Central Dogma of Biology: DNA --> RNA --> Protein

DNA is the carrier of genetic information in organisms. What does that mean? Large molecules in organism can have many functions: they can provide structure, act as catalyst for chemical reactions, serve to sense changes in their environment (leading to immune responses to foreign invaders and to neural responses to stimuli such as light, heat, sound, touch, etc) and provide motility. DNA really does none of these things. Rather you can view it as an information storage system. The information must be decoded to allow the construction of other large molecules. The other molecules are usually proteins, another class of large polymers in the body. Chromosomes are located in the nucleus of a cell.

- **Great animations of DNA and the Central Dogma of Biology** from the HHMI.

Replication

Chromosomes are located in the nucleus of a cell. DNA must be duplicated in a process called **replication** before a cell divides. The replication of DNA allows each daughter cell to contain a full complement of chromosomes.

- **DNA Anatomy: Animations of Structure, Replication, Transcription, and Translation**
- **Animation of Replication**

Transcription

The actual information in the DNA of chromosomes is decoded in a process called **transcription** through the formation of another nucleic acid, ribonucleic acid or RNA. The RNA, made by the enzyme RNA polymerase, is complementary to one strand of the DNA. RNA differs from DNA in that RNA contains a ribose, not deoxyribose, sugar in its backbone. In addition, RNA lacks the base T. It is replaced, instead, with the base U, which is complementary to A (as T is complementary to A in DNA). The RNA formed acts as a messenger, which passes from the nucleus into the cytoplasm of the cell. In fact, this type of RNA is often called messenger RNA, mRNA. Since the information in a nucleic acid (DNA) is converted into information in the form of another nucleic acid (RNA), this process is called transcription (since the language is still the same, such as when you transcribe a written speech in English into written English).

- **animation of transcription**

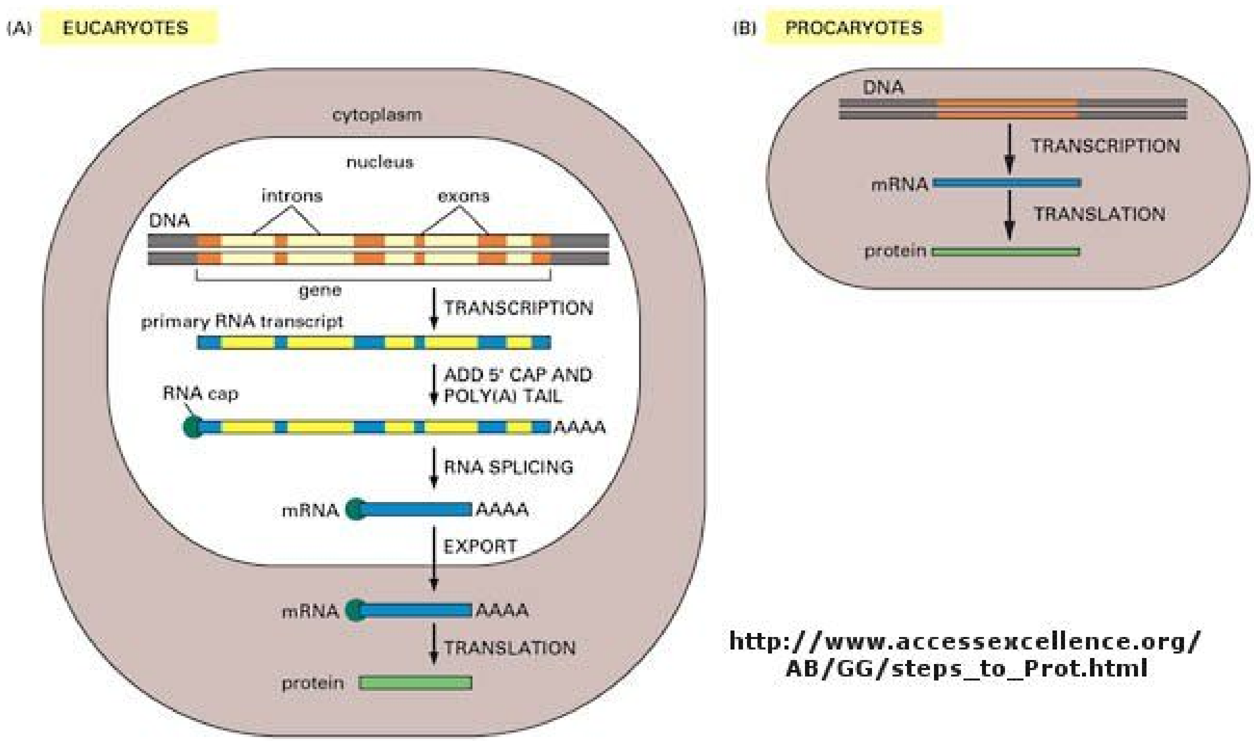
Translation

The information from the DNA, now in the form of a linear RNA sequence, is decoded in a process called **translation**, to form a protein, another biological polymer. The monomer in a protein is called an amino acid, a completely different kind of molecule than a nucleotide. There are twenty different naturally occurring amino acids that differ in one of the 4 groups connected to the central carbon. In an amino acid, the central (alpha) carbon has an amine group (RNH_2), a carboxylic acid group (RCO_2H), and H, and an R group attached to it. Since the information in a nucleic acid (RNA) is converted into information in the form of a different molecule, a protein, this process is called **translation** (since the language of nucleic acids is changed to that of proteins, such as when you translate English into Chinese).

In contrast to the complementarity of DNA and RNA (1 base in RNA complementary to 1 base in DNA), there is not a 1:1 correspondence between a base (part of the monomeric unit of RNA) in RNA to the monomer in a protein. After much work it was discovered that a contiguous linear sequence of 3 nucleotides in RNA is decoded by the molecular machinery of the cytoplasm with the result that 1 amino acid is added to the growing protein. Hence a triplet of nucleotides in DNA and RNA have the information for 1 amino acid in a protein. That there was not a 1:1 correspondence between nucleotides in nucleic acids and amino acids in proteins was evident long ago since there are only 4 different DNA monomers (with A, T, G, and C) and 4 different RNA monomers (with A, U, G, and C) but there are 20 different amino acid monomers that compose proteins.

- **animation: translation**

Now, it turns out that not all the information in the DNA sequence of a organism encodes for a protein. In fact only about 2% of the 3 billion base pairs seem to be transcribed into RNA which can be translated into protein. The function of the rest of the DNA is at present uncertain. How does the molecular machinery of the cell know which part of the DNA encodes for proteins. It turns out that there are unique DNA sequences at the beginning and end of the part of the DNA sequence that codes for a protein. Proceed down the DNA of a chromosome and suddenly you come to those signals, which are recognized by the cells machinery. A complementary RNA is made from that section, and the complementary RNA is then decoded into a single protein. Continue further down the DNA sequence and another such coding sequence is found, which can be transcribed into a mRNA, which then can be translated into another unique protein. In all there are about 30,000 such sections of DNA in all the chromosomes that encode the information for 30,000 unique proteins. These unique coding sections of DNA that ultimately are transcribed into unique mRNA which are translated into unique proteins are called **genes**. For our purposes, we conclude that one gene has the information for one protein. Each of the protein differ from each other in both length, and the specific sequence of amino acids in the protein. The DNA is indeed the blueprint of the cell. What determines the actual characteristics of the cells are the actual proteins that are made by the cell.



http://www.accessexcellence.org/AB/GG/steps_to_Prot.html

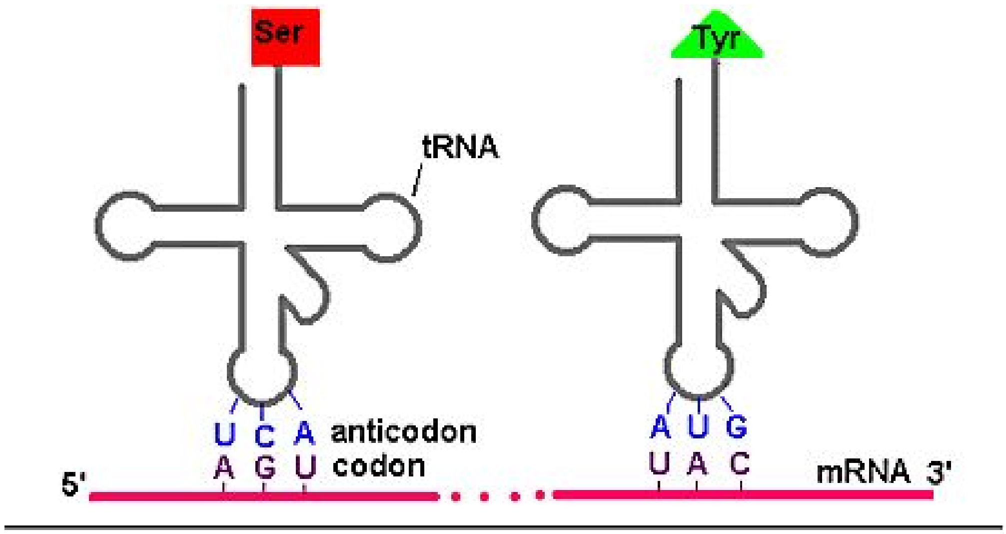
Not only must DNA be transcribed into DNA, but the genetic information in the DNA must be replicated before a given cell divides, so that the daughter cells both contain the same genetic information. In replication, the dsDNA separate, and an enzyme, DNA polymerase, makes complimentary copies of each strand. The two resulting dsDNA strands separate to different daughter cells during division. The process whereby DNA is replicated when cells divide, and is transcribed into RNA which is translated into protein is called the **Central Dogma of Biology**.

As mentioned above, each amino acid is specified by a particular combination of three nucleotides in RNA. The three bases are called a codon. The Genetic Code consists of a chart which shows what triplet RNA sequence or codon in mRNA codes for which of the 20 amino acids. One of the codon codes for no amino acids and serves to stop the synthesis of the protein from the mRNA sequence. The genetic code is shown below:

GENETIC CODE

<table>
<tr><th>FIRST POSITION</th><th colspan="4">SECOND POSITION</th><th>THIRD POSITION</th></tr>
<tr><th></th><th>U</th><th>C</th><th>A</th><th>G</th><th></th></tr>
<tr><td rowspan="4">U</td><td rowspan="2">phenyl-alanine</td><td rowspan="4">serine</td><td rowspan="2">tyrosine</td><td rowspan="2">cysteine</td><td>U</td></tr>
<tr><td>C</td></tr>
<tr><td rowspan="2">leucine</td><td>stop</td><td>stop</td><td>A</td></tr>
<tr><td>stop</td><td>tryptophan</td><td>G</td></tr>
<tr><td rowspan="4">C</td><td rowspan="4">leucine</td><td rowspan="4">proline</td><td rowspan="2">histidine</td><td rowspan="4">arginine</td><td>U</td></tr>
<tr><td>C</td></tr>
<tr><td rowspan="2">glutamine</td><td>A</td></tr>
<tr><td>G</td></tr>
<tr><td rowspan="4">A</td><td rowspan="3">isoleucine</td><td rowspan="4">threonine</td><td rowspan="2">asparagine</td><td rowspan="2">serine</td><td>U</td></tr>
<tr><td>C</td></tr>
<tr><td rowspan="2">lysine</td><td rowspan="2">arginine</td><td>A</td></tr>
<tr><td>* methionine</td><td>G</td></tr>
<tr><td rowspan="4">G</td><td rowspan="4">valine</td><td rowspan="4">alanine</td><td rowspan="2">aspartic acid</td><td rowspan="4">glycine</td><td>U</td></tr>
<tr><td>C</td></tr>
<tr><td rowspan="2">glutamic acid</td><td>A</td></tr>
<tr><td>G</td></tr>
</table>

* and start



2nd base in codon

1st base in codon	U	C	A	G	3rd base in codon
U	Phe Phe Leu Leu	Ser Ser Ser Ser	Tyr Tyr STOP STOP	Cys Cys STOP Trp	U C A G
C	Leu Leu Leu Leu	Pro Pro Pro Pro	His His Gln Gln	Arg Arg Arg Arg	U C A G
A	Ile Ile Ile Met	Thr Thr Thr Thr	Asn Asn Lys Lys	Ser Ser Arg Arg	U C A G
G	Val Val Val Val	Ala Ala Ala Ala	Asp Asp Glu Glu	Gly Gly Gly Gly	U C A G

The Genetic Code www.accessexcellence.org/AB/GG/genetic.html

Determining the protein sequence from a DNA sequence.

For a given gene, only one strand of the DNA serves as the template for transcription. An example is shown below. The bottom (blue) strand in this example is the **template** strand, which is also called the **minus (-)** strand, or the **sense** strand. It is this strand that serves as a template for the mRNA synthesis. The enzyme RNA polymerase synthesizes an mRNA in the 5' to 3' direction complementary to this template strand. The opposite DNA strand (red) is called the **coding** strand, the **nontemplate** strand, the **plus (+)**strand, or the **antisense** strand.

The easiest way to find the corresponding **mRNA sequence** (shown in green below) is to read the **coding, nontemplate, plus (+)**, or **antisense** strand directly in the 5' to 3' direction substituting U for T. Find the triplet in the coding strand, change any T's to U's, and read from the Genetic Code the corresponding amino acid that would be incorporated into the growing protein. (The strands below are separated into triplets for ease of visualization.) An example is show below.

DNA 5' A G G C C T T C G A A C G G G A T G G A A T G A 3'
DNA 3' T C C G G A A G C T T G C C C T A C C T T A C T 5'
RNA 5' A G G C C U U C G A A C G G G A U G G A A U G A 3'
PROT N ARG PRO SER ASP GLY MET GLU STOP

- **DNA or RNA sequence to Amino Acid converter**
- **reverse translate a protein into a DNA sequence** | **Another reverse translator**