# MODULE 1: Molecular Basis of Inheritance

## Learning Outcomes:

- Describe the structure of DNA including the structure of nucleotides and bases.
- Describe why DNA strands are antiparallel and complementary and define 5' and 3' ends.
- Describe DNA replication including direction of synthesis of DNA strands, enzymes involved and understand why Okazaki fragments exist.
- Describe the processes of transcription and translation and the flow of genetic information from DNA to protein.
- Describe the differences between DNA and RNA.
- Describe the common consensus sequences found in promoters of prokaryotic and eukaryotic genes including position and function.
- Describe the genetic code and remember the sequence of start and stop codons.
- Describe the general structure and interactions of the ribosome, aminoacyl-tRNA and mRNA during translation.
- Describe the intron /exon structure of eukaryotic genes and understand why introns are removed from primary transcripts before translation can occur.
- Describe the causes of and types of mutations and define the terms 'missense mutation', 'nonsense mutation', 'silent mutation', 'point mutation', 'substitution mutation' and 'frameshift mutation'.

## Introduction - What is DNA

All living organisms are made of up cells. Some life forms consist of only a single cell (e.g. bacteria), while others are made up of trillions of cells (e.g. humans or animals). Just to give you some idea, the average human heart is made of up ~2 billion heart muscle cells; and the average human hand will contain ~2.5 billion cells.

Multicellular organism are also made up of many different types of cells. For example, retinal cells (those that make up the retina in your eyes) enable us to detect light and see; red blood cells carry oxygen from lungs to different parts of the body; white blood cells allow us to fight infectious organisms that invade our bodies; intestinal cells help us to digest and absorb food etc etc. Therefore, you can think of each cell type having a specific function. Since our bodies are made up of more than 200 different types of cells, it is safe to assume that these cells perform at least 200 different functions.

So, the natural question is - how is the function of any given cell type determined. The answer is simple - cells are told what they are to do by a key component of each cell called '**DNA**'. The goal of this unit is to understand what makes up DNA and how DNA instructs cells to carry out specific functions.

## The Structure of Deoxyribonucleic Acid (DNA)

The genetic material of a cell consists of a mixture of protein and DNA. DNA is the component of interest because it contains the genes and information required for controlling cellular function and heredity. DNA is a long, threadlike polymer or macromolecule which consists of monomer units called **deoxyribonucleotides**.

Deoxyribonucleotides consist of three components:

- nitrogenous base
- sugar
- phosphate group(s)

The **sugar component** is deoxyribose (hence the name **D**eoxyribo-**N**ucleic **A**cid - or **DNA**) which is a derivative of ribose. Deoxyribose and ribose are carbohydrates that are similar in structure (Figure 1.1a), except that **deoxyribose** lacks an oxygen atom attached to carbon # 2 in the deoxyribose molecule.

The **nitrogenous** bases are nitrogen containing cyclic compounds with the purines having double ring and the pyrimidines having a single ring structures (Figure 1.1a). The two **purine** bases are **adenine** and **guanine** and the two **pyrimidines** are **cytosine** and **thymine**. In a double stranded DNA molecule, hydrogen bonds between these nitrogenous bases on opposite strands is what keeps the strands together. Hydrogen bonding is specific and Adenine (**A**) always pairs with Thymine (**T**) via two hydrogen bonds; and Guanine (**G**) always pairs with Cytosine (**C**) via three hydrogen bonds.

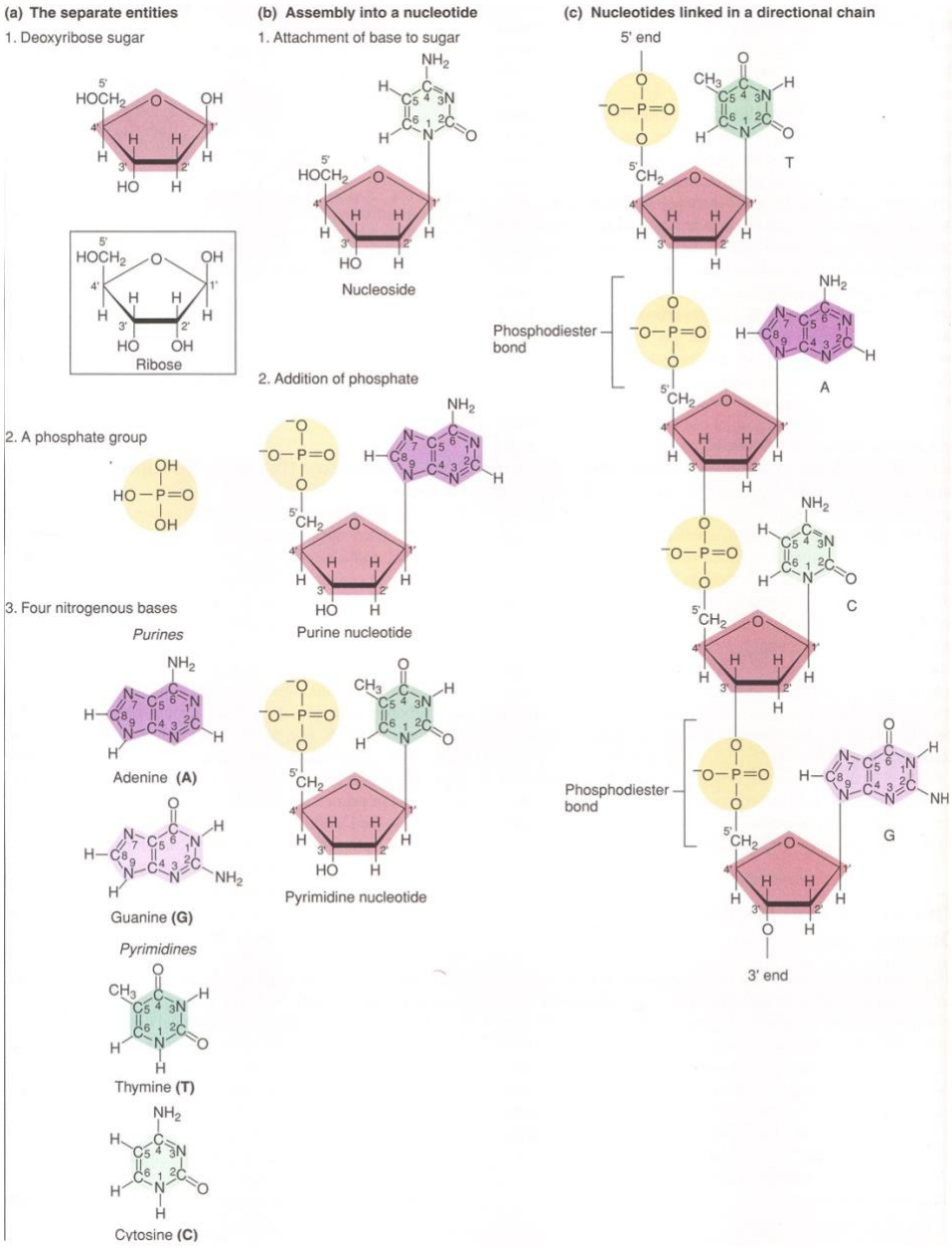


***Figure 1.1:*** *Chemical structures of the components of DNA. The structures of the sugar, phosphate and base components are shown in (a) along with nucleotides 'monomers' in (b) and a single strand of DNA 'polymer' in (c).*
*Source: Hartwell et al. 2006, Genetics: From genes to genomes, p. 174.*

A **nucleotide** is composed of a phosphate group attached to carbon 5 of deoxyribose along with the base attached to carbon 1. Nucleotides containing three phosphate groups (tri-phosphates) are the precursors in the synthesis of DNA. However during DNA synthesis two of the phosphate groups are lost so that the nucleotides in the DNA molecule contain only one phosphate group. Nucleotides therefore vary only in the bases they have being either A, C, G or T.

A DNA molecule consists of a backbone which is invariant. The backbone is made of the alternate deoxyribose sugar and phosphate groups. Two nucleotides are joined together via the phosphate group which is attached to carbon 3 (3') and carbon 5 (5') of the deoxyribose sugars of two neighbouring nucleotides. The fact that the two ends of the DNA molecule are not the same (i.e. the 5'-end contains a phosphate group and the 3'-end contains a hydroxyl (OH) group) gives the molecule **polarity**. Polarity meaning that the molecule is asymmetrical.

The order of the bases in DNA is fundamentally important in defining the genetic code. Every gene contains a unique sequence of bases. The structure of one strand of DNA is shown in Figure 1.1c.

## The Watson-Crick Double Helix

In cells DNA chains exist as double stranded molecules, that is, two DNA chains wound around one another. The structure of the two chains is called the DNA double helix as shown in Figure 1.2.

This model was first suggested by James Watson and Francis Crick in 1953. Watson and Crick along with Maurice Wilkins were awarded the Nobel Prize for this discovery. Rosalind Franklin may have also received the Nobel Prize had she not died of cancer prior to the award. Nobel prizes are not given posthumously! Bugger!

The five important features of the DNA double helix are:

1. Two helical polynucleotide chains are coiled around a common axis. The two DNA chains run in opposite directions, i.e. the 5' end of one chain is next to the 3' end of the other chain. They are **antiparallel**.
2. The purine and pyrimidine bases are on the inside of the helix, whereas the phosphate and deoxyribose units are on the outside. The planes of the bases are perpendicular to the helix axis. The planes of the sugars are at right angles to those of the bases.
3. The helical structure repeats after ten nucleotide residues. Therefore nucleotides that are ten residues apart are on the same side of the helix.
4. The two chains are held together by hydrogen bonds between the bases that face each other on the inside of the helix (termed **base pairs**). Adenine can only base pair with thymine. Guanine can only base pair with cytosine. A-T base pairs consist of two hydrogen bonds whereas G-C base pairs consist of three hydrogen bonds. Consequently G-C base pairs are stronger that A-T base pairs.
5. The sequence of bases along the DNA chain is not restricted in any way. The precise sequence of bases carries the genetic information.

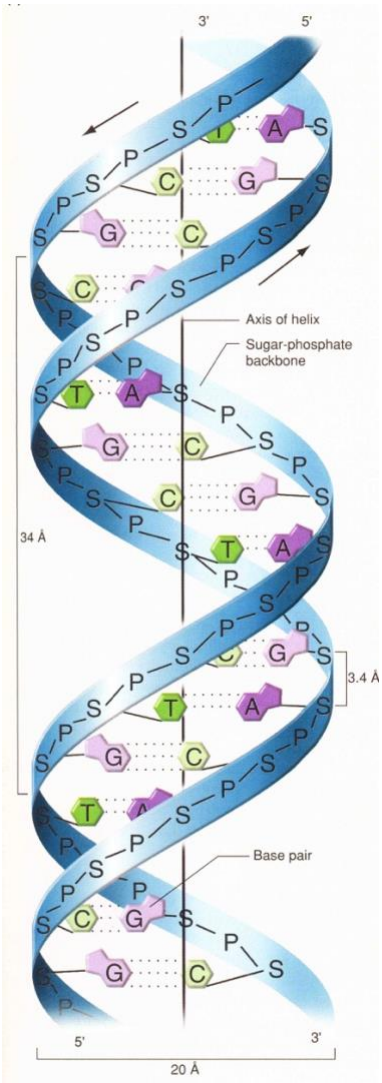



*Figure 1.2: Ribbon model of the double helix of DNA.*
*Horizontal dotted lines represent hydrogen bonds between bases. S = sugar and P = phosphate groups.*
*Source:* Hartwell *et al.* 2006, *Genetics: From genes to genomes,* p. 178.

**Complementary nature of the two DNA chains in the double helix**

As mentioned above the two chains of the helix run in opposite directions. If drawn diagrammatically it looks like this:

5' A G C T T G C A T 3'
3' T C G A A C G T A 5'

Because A always pairs with T and G always pairs with C, the order of the bases on one strand dictates the order of bases on the other strand. Hence one chain is the **complement** of the other. It is important to realise that the two chains are NOT the same. For example lets look at the sequence of the above from the 5' end of each strand:

top strand 5' A G C T T G C A T
bottom strand 5' A T G C A A G C T

They complement each other but they are not the same.

## DNA Replication

The complimentary nature of the two DNA strands is the key to the way in which DNA is replicated. During DNA replication the two DNA strands are separated by the breaking of the hydrogen bonds between the bases. Each strand then acts as a template for the synthesis of a new DNA strand, the nature of the new DNA strand being determined by the order of the bases on the template strand. This is termed **semiconservative** replication and is demonstrated in Figure 1.3.

Each strand of the double helix is used as a template to synthesise a new strand. The process of DNA replication results in two molecules of DNA with each composed of one old strand and one new strand.

DNA replication involves a number of different enzymes. The enzyme that synthesises the new DNA strand is called **DNA polymerase**. DNA polymerase catalyzes the step-by-step addition of deoxyribonucleotide units to a DNA chain. Deoxyribonucleotide triphosphates (dNTPs) are used as the source of nucleotides, i.e. dATP, dCTP, dGTP and dTTP. DNA polymerase adds a deoxyribonucleotide to the 3' OH group of the newly synthesised strand. **DNA is therefore synthesised only in the 5' to 3' direction** and the incoming dNTP is always added to the 3' OH group of the preceding nucleotide (deoxyribose).

This new deoxyribonucleotide is determined by the opposite base on the template strand, i.e. if the template has a G the new incoming deoxyribonucleotide must have the base C or cytosine (deoxycytidine). The dNTP loses two phosphate groups, because the phosphodiester bond involves only one phosphate group.

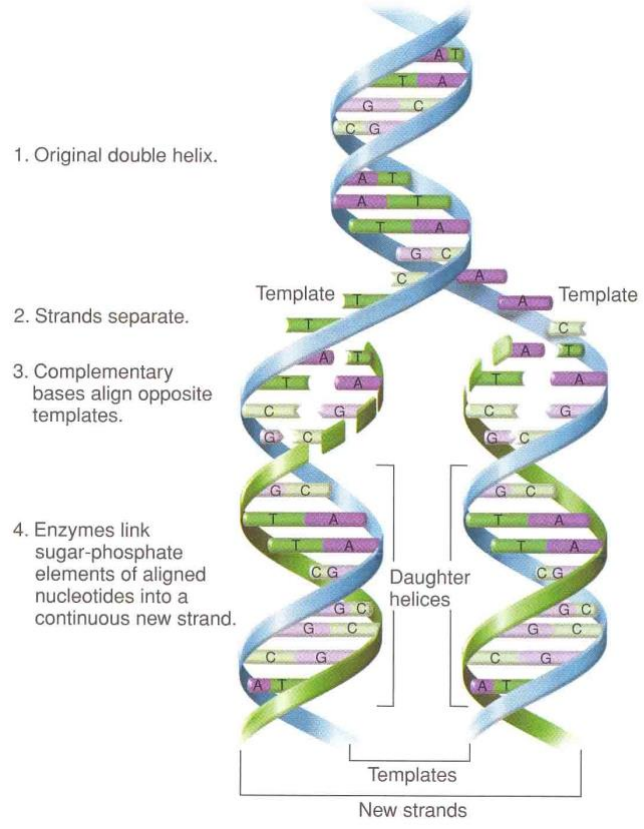


***Figure 1.3:*** *Semiconservative replication of DNA.*
*Source: Hartwell et al. 2006, Genetics: From genes to genomes, p. 185*

**DNA Replication Summary**

The first step in DNA replication is to 'unzip' the double helix structure of the DNA molecule. This is carried out by an enzyme called ***helicase*** which breaks the hydrogen bonds holding the complementary bases of DNA together (A with T, C with G).

The separation of the two single strands of DNA creates a 'Y' shape called a ***replication fork***. The two separated strands will act as templates for making the new strands of DNA. One of the strands is oriented in the 3' to 5' direction (towards the replication fork), this is the ***leading strand***. The other strand is oriented in the 5' to 3' direction (away from the replication fork), this is the ***lagging strand***.

As a result of their different orientations, the two strands are replicated differently:

**Leading Strand:** A short piece of RNA called a ***primer*** (produced by an enzyme called ***primase***) comes along and binds to the end of the leading strand. The primer acts as the starting point for DNA synthesis. ***DNA polymerase*** binds to the leading strand and then walks along it, adding new complementary nucleotide bases (A, C, G and T) to the strand of DNA in the 5' to 3' direction. This sort of replication is called continuous.

**Lagging Strand:** Numerous RNA primers are made by the primase enzyme and bind at various points along the lagging strand. Chunks of DNA, called ***Okazaki fragments***, are then added to the lagging strand also in the 5' to 3' direction. This type of replication is called ***discontinuous*** as the Okazaki fragments will need to be joined up later.

**Size of DNA molecules**

DNA molecules are very, very long. Mammalian genomes all have approximately $3 \times 10^9$ base pairs or nucleotide pairs (3000 million base pairs). One can understand the necessity for this in considering that all the different proteins of an organism must be encoded by the organisms DNA. There are thousands of different proteins even in a bacterium so DNA molecules are a very long. Humans have 46 chromosomes and cows have 60. The smallest human chromosome is chromosome 21 and it has 47 million base pairs where the largest , chromosome 1 has 247 million base pairs by comparison. If all the chromosomes were laid end to end the length of DNA would be approximately 2 meters which is 8000 times the length of a nucleus. DNA is therefore normally in a condensed form, within the cell, tightly coiled and compacted with the help of proteins, one class of which is called histones.

**Some viral genomes are made of RNA**

The vast majority of organisms have a double stranded DNA genome. However viruses show a considerable degree of variation in the form of their genome. Viral genomes may be single or double stranded DNA and some viruses have a genome consisting of RNA which again may be single or double stranded and **positive or negative sense** which refers to which strand is functional (we will come to this later).

## Flow of Information from DNA to Protein Synthesis

The sequence of the bases in a DNA molecule specifies genetic information. The molecules which perform most biological reactions within cells and form all the structural components are

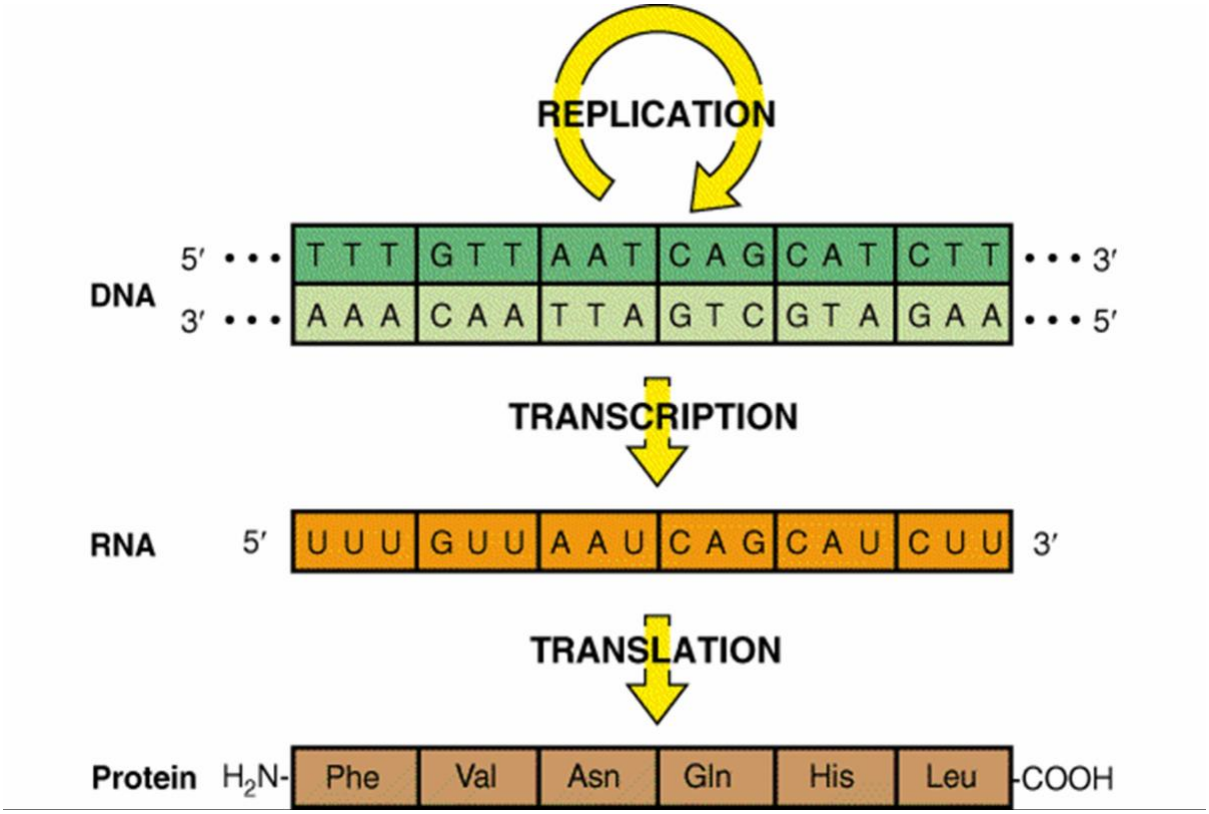
proteins. It is arguable that it is proteins rather than DNA which are most important to the cell. All proteins are made up of strings of monomer units called amino acids. There are 20 different amino acids that can be put into protein and it is the number and specific order of these that specifies the proteins function. The order is specified by the bases in DNA. More specifically the base sequence corresponds to the amino acid sequence of a protein. Proteins perform a very wide variety of functions. So the DNA of a cell determines the proteins present in that cell.

When we portray the genetic material as being the blueprint or major controlling centre of the cell, it is through its function in determining what proteins the cell produces. These proteins in turn determine the chemical reactions of the body.

**How genes specify the production of a proteins**

DNA is not the template for protein synthesis, rather the template for protein synthesis is RNA.

The flow of genetic information is as follows:




Essentially a gene is a stretch of DNA a few basepairs to several thousand in length. A copy of this is made in the form of a messengerRNA (mRNA). This initial stage is called **TRANSCRIPTION**. This mRNA then serves as a template for synthesis of a protein. This occurs in the cytoplasm of a cell on specialised organelles called ribosomes. Ribosomes essentially translate the code in the mRNA from a series of bases to a series of amino acids and hence this process is called **TRANSLATION**. The easiest way to get to grips with the details of these processes to study them separately and to then consider how they are controlled and integrated to produce the many thousands of different proteins within the cell. The process of going from a gene to a protein is called **'gene expression'** and is also often defined as **protein synthesis**, however protein synthesis is strictly the same as translation.

## RNA

In this topic we are going to study the three types of RNA involved in the process of going from DNA to protein synthesis; also how the sequence of bases in DNA specifies the amino acid sequence of a protein (**genetic code**). Then we are going to look at the structure of genes.

DNA is short for **d**eoxyribo**n**ucleic **a**cid being composed of deoxyribonucleotides. RNA on the other hand is short for **r**ibo**n**ucleic **a**cid being composed of ribonucleotides. Hence RNA molecules, like DNA, is a long polymers of ribonucleotides.

Ribonucleotides consist of the same three components as DNA, namely:

1. sugar
2. nitrogenous base
3. phosphate group(s)

There are **three fundamental differences between DNA and RNA**:

1. The sugar in RNA is ribose, as opposed to deoxyribose in DNA.
2. RNA contains four different types of nitrogenous bases; three are the same as DNA, i.e. adenine, guanine and cytosine. The fourth is uracil (instead of thymine). Like thymine, uracil is a pyrimidine.
3. Unlike DNA, RNA is a single stranded molecule, except in some viruses. However, RNA commonly forms double stranded regions.

RNA follows the same base pairing rules as DNA, i.e.

- guanine pairs with cytosine G-C
- adenine pairs with uracil A-U.

RNA molecules have a wide variation in size, from as few as 75 bases to many thousands.

### Three types of RNA involved in protein synthesis

There are several different types of RNA in the cell. Three different types are involved in protein synthesis. These are:

- ribosomal RNA - rRNA
- transfer RNA - tRNA
- messenger RNA - mRNA

**Messenger RNA** (mRNA) is the template for protein synthesis. DNA is copied into mRNA which carries the information of the DNA to the site of protein synthesis, the ribosomes. Every gene or group of genes has a corresponding mRNA molecule. Therefore, the size of mRNAs varies considerably.

**Transfer RNA** (tRNA) carries amino acids to the ribosomes for incorporation into proteins. The order of the amino acids is determined by the mRNA template. There is at least one tRNA for each of the 20 amino acids. Transfer RNA molecules are only 75 nucleotides long, being the smallest of the RNA molecules.

**Ribosomal RNA** (rRNA) is the major component of ribosomes. There are several types of ribosomal RNA that vary in size. rRNA is the most abundant of all RNA constituting some 80% of total RNA.

## Transcription and Translation

### Transcription and structural genes

Messenger RNA (mRNA) is synthesised in a similar way to DNA. The enzyme responsible is **RNA polymerase**. The precursors are ribonucleoside triphosphates (NTPs). RNA synthesis occurs in the 5' to 3' direction. RNA polymerase like DNA polymerase, catalyses the formation of the phosphodiester bond between nucleotides. Again this only occurs if the incoming nucleotide contains a base that is complementary to that on the DNA template strand. The synthesis of mRNA is conservative, unlike DNA replication which is semi-conservative. The DNA template is conserved, i.e. left in original condition. The DNA remains intact while many mRNA copies are made. During mRNA synthesis the two strands of the DNA are unwound; only one strand acts as the template. Consequently the mRNA sequence is the same as one DNA strand (coding strand) and the complement of the other strand (template or non-coding).

Hence it can be seen that mRNA copies the code of the bases from DNA. The mRNA molecule then moves to the ribosomes where translation (protein synthesis) takes place.

Transfer RNA (tRNA)and rRNA molecules are also made by copying sequences present in DNA. Therefore DNA contains both genes encoding proteins and sequences corresponding to tRNA and rRNA. The function of the tRNA and rRNA sequences in DNA is solely for the production of these RNAs.

Genes are discrete sections of DNA that contain the sequence corresponding to a protein. In single cell organisms (procaryotes), like bacteria, several genes are copied into a single RNA molecule. A cluster of genes copied by a single RNA is called an **operon.** In higher organisms, like plants and animals, each gene is transcribed into a separate mRNA molecule.

### Promoters and terminators

Special sequences in DNA specify the start and stop positions for transcription. These sequences are recognised by RNA polymerase and initiate and terminate the synthesis of mRNA.

The -10 and -35 regions refer to sequences that are present in the upstream region of the gene. These sequences are bound by RNA polymerase and promote the initiation of mRNA synthesis from the +1 nucleotide. These regions are called **promoter sites**. -10 and -35 refer to the number of nucleotides away from the initiation site. The -10 or pribnow box has the sequence TATAAT. The -35 region has the sequence TTGACA.

To terminate transcription most genes have 'terminator' sequences which prevent RNA polymerase transcribing any further. Asta La Vista baby! It is the RNA polymerase that says "I'll be back" in this film!

In higher organisms (eucaryotes) the transcription initiation and termination signals are slightly different from prokaryotes. The analogous consensus promoter sequences are -25 (TATA box) and -75 (CAAT box) bases upstream of the initiation site of mRNA. These are called the TATA and CAAT boxes, respectively

**Translation**

Once mRNA is synthesised it moves to the ribosomes. A ribosome attaches at the 5' end and tracks along the mRNA until it comes to the initiation codon which has the sequence AUG. A process then occurs whereby the sequence of bases on mRNA is read and utilised to construct proteins from the pool of amino acids. Transfer RNA is an **adaptor** molecule that recognises the base sequence of mRNA and carries the corresponding amino acid in readiness for protein synthesis to the ribosome. Transfer RNAs contain both an amino acid attachment site and a template (mRNA)-recognition site.

A tRNA molecule carries a specific amino acid to the site of protein synthesis. The template recognition site on the tRNA molecule is a sequence of three bases called the **anticodon.** These three bases are complementary to three bases on mRNA called the codon (see below). It is this interaction which hold the tRNA at the ribosome in position so that the amino acid it carries can be covalently linked to the next amino acid brought in on another tRNA, as the ribosome moves along the mRNA toward the 3' end. The ribosome covalently links the amino acids by 'peptides bonds' and detaches them from the tRNA molecules which are then released from the ribosome. Only the correct tRNA with a complementary anticodon base pairs with the codon such that the correct amino acid is added to the growing protein chain. Having read this now look at Figure 1.5 summarising translation. Can you see what is wrong with it? See Figure legend for answer.

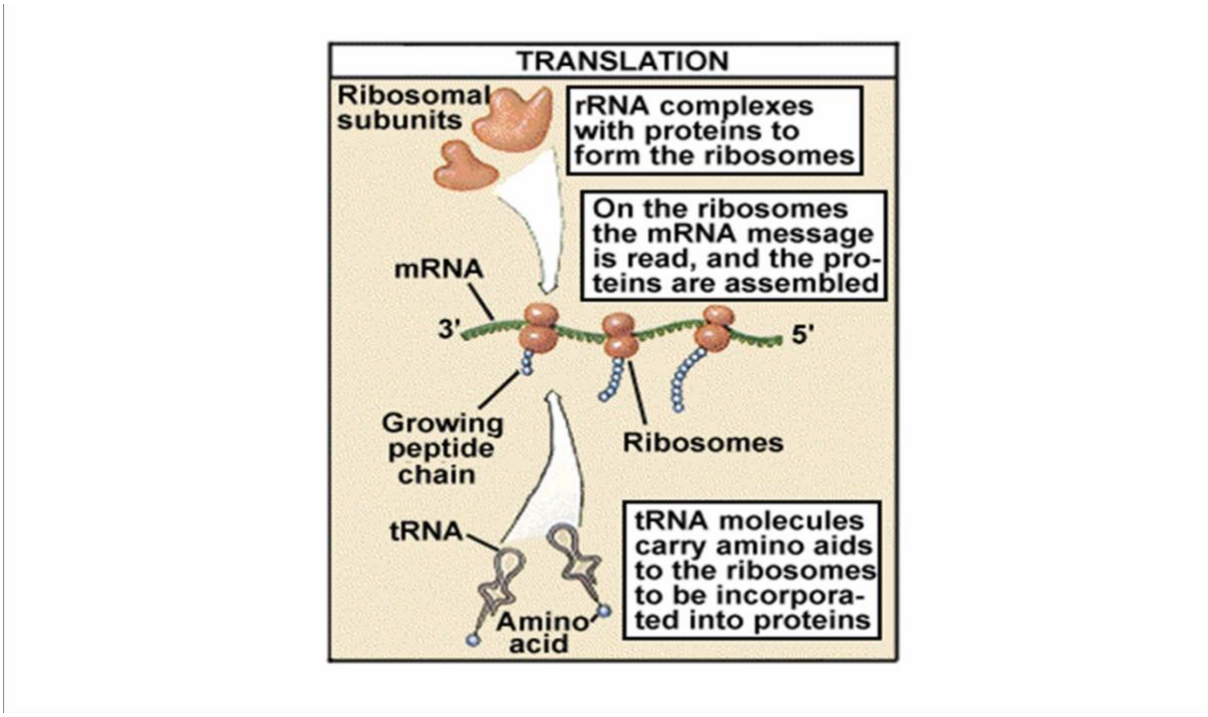



***Figure 1.5:*** *Summary of translation.*
*The artists error here is that the orientation of the mRNA molecule should read 5' to 3'.*
*Source: Black, J 2002, Microbiology principles and exploration, 5th edn, p. 173.*

## The Genetic Code

The genetic code is the relationship between the sequences of bases in DNA (or the mRNA transcript) and the sequence of amino acids in a protein. Each amino acid is coded by a group of three bases called a codon.

The genetic code is non-overlapping and begins from a fixed point.

e.g. sequence of DNA ..... ACT/GTC/TAC/GTA/CTT/.... Protein amino acid sequence ..... aa1 aa2 aa3 aa4 aa5 .....

A deletion of one base changes the amino acid sequence of the protein.

e.g. ACT/GTC/ACG/TAC/TTA/ ....

aa1 aa2 aax aay aaz .....

Similarly the addition of a base changes the amino acid sequence of the protein.

e.g. ACT/GTC/TTA/CGT/ACT/TAC/.....

aa1 aa2 aac aad aae aaf .....

**Major features of the genetic code**

If three bases encode one amino acid, and there are four different bases, then there are 43 = 64 different possible codons. There are twenty amino acids therefore each amino acid must be coded by more than one codon. Indeed this is the case. Several series of experiments between 1961 and 1966 deciphered all 64 of the codons in terms of which amino acid they correspond to. The full genetic code is found in your textbook figure 14-7, p.360. Sixty one of the triplets code for particular amino acids and three code for chain termination, i.e. these codons signal the end of the protein.

**Degeneracy of the genetic code**

One thing we notice immediately about the genetic code is that many amino acids are coded by more than one codon. Only methionine (Met) and tryptophan (Trp) are coded by a single codon. Some amino acids are coded by six codons, e.g. leucine (Leu), arginine (Arg) and serine (Ser).

This is referred to as **degeneracy** of the genetic code. It functions to minimise the deleterious effects of mutations by allowing any base to exist in the third position of the codon, whilst still coding for the same amino acid.

The genetic code is universal, i.e. it is the same from bacteria (procaryotes) to humans (eucaryotes), except for some minor exceptions in one or two organisms.

**Translation start and stop signals**

Certain codons do not encode specific amino acids, but rather are signals indicating the end of the protein coding sequence. These codons are: UAA, UAG, and UGA and are termed **stop codons**. Similarly the codon AUG indicates the beginning of the protein coding sequence and is termed a **start codon**. AUG also codes for methionine (Met).

**The sequences of genes and their encoded proteins are colinear**

Genes and their protein products are collinear. This means the sequence of nucleotides in DNA from the 5' to 3' end of the gene and mRNA corresponds to the sequence of amino acids in the

protein product from the amino-terminus to the carboxy-terminus, respectively. See section 15.8, p.396 of your text. Note the protein or polypeptide is made starting from the N-terminus to the C-teminus.

**Eucaryotic genes – exons and introns**

The genes of higher organisms (eucaryotes) are not continuous but are interrupted by non-coding intervening sequences called **introns.** An extreme example is shown is the dystrophin gene. People and animals suffering from muscular dystrophy have a muscle wasting disease which is caused by mutations in the gene encoding this muscle protein.

The dystrophin gene is massive and weighs in at 2.4 megabasepairs (2400 kb). The mRNA it produces is only 14 kb long and the gene contains over 80 introns which are removed to produce the mature mRNA.

The regions of the gene containing sequences corresponding to the amino acid sequence of the protein are called **exons** (expressed regions). These genes are transcribed into a long mRNA molecule. The intron sequences are spliced out of the mRNA so that only the exons remain joined together in one continuous sequence. This spliced mature mRNA is then used as a template for protein synthesis. The details of this process are not important for you. Introns do not occur in the genes of procaryotes.

## Mutations

Mutations are changes in the sequence of base pairs in DNA. These can be large deletions or insertions of many base pairs of DNA or they can be single base deletions insertions or substitutions. They can add or remove genes from a genome or they can cause abberant expression of genes. They do not all occur in the coding sequences of genes and you should think about how the expression of a gene might be altered if they occur in non-coding DNA or promoters. Will gene expression it be altered at all? Mutations are a major source of genetic variation and can lead to genetic disease or new traits in all organisms. Mutations can be caused by mistakes in DNA replication and repair, uptake and insertion of DNA by an organism (naturally or genetically engineered), chemical mutagens and different sorts of radiation we a constantly exposed to e.g. UV, X-rays and gamma rays.