Some Basic Molecular Biology

This is intended to be a very brief introduction to molecular biology, especially to DNA, RNA and protein sequences.
Inheritance

One of the basic problems of biology is to understand inheritance.

In 1865 Mendel gave an abstract, essentially mathematical model of inheritance in which the basic unit of inheritance was a gene.

Mendel’s work was forgotten until 1900, and then it was taken up again and underwent intense mathematical development.

Still the nature of the gene was unknown.

In 1944 the gene was known to be made of DNA.

It was until 1953 that James Watson and Francis Crick proposed the now famous double helical structure of DNA.

This model describes the gene itself, providing the basis for a deeper understanding of inheritance.
The molecules of the cell are of two classes: large and small.

The large molecules, known as macromolecules, are of three types: DNA, RNA and protein.

These macromolecules are made by joining certain small molecules together in polymers.
DNA is the basis of inheritance. It is a nucleic acid and is made up of small molecules called nucleotides.

There are four different nucleotides: adenine (A), Cytosine (C), Guanine (G), and thymine (T).

RNA is also a nucleic acid and is made up of ribonucleotides: adenine (A), Cytosine (C), Guanine (G), and Uracil (U).

DNA and RNA both have directionality, one end is called 5’ and the other 3’.

Proteins are also polymers and have an alphabet of 20 amino acids. Proteins also have directionality.
DNA

The DNA contained in the cell is called the genome.

How much DNA does an organism need to function?

The genome of the intestinal bacterium Escherichia coli (E. coli), an organism with one cell, has about $5 \times 10^6$ letters.

The human genome is about $3 \times 10^9$ letters. Each human cell contains the same DNA.
The key feature of DNA that suggested the copying mechanism is the complementary base pairs.

A is paired with T and G is paired with C. These pairings are by hydrogen bonds. The three-dimensional structure is helical. It is shown as double helix of backbone with bases, or letters, attached.

DNA carries the genetic material, i.e. the information required by an organism to function.

There are exceptions: for certain viruses, the genetic material is RNA.

DNA is also the means by which organisms transfer genetic information to their descendants.
The central dogma

In eukaryotes, organisms with a nucleus, DNA remains in the nucleus; whereas protein are made in the cytoplasm outside the nucleus.

RNA is the intermediate molecule which carries the information out of nucleus.

The information flow in biology is summarized by the "central dogma" put forward by Francis Crick in 1958:

The central dogma states that once "information" has passed into protein it cannot get out again. The transfer of information from nucleic acid to nucleic acid, or from nucleic acid to protein, may be possible, but transfer from protein to protein, or from protein to nucleic acid is impossible.

Information here means the precise determination of sequence, nucleotides in DNA, ribonucleotides in RNA, or amino acid residues in the protein.
The general idea is that one macromolecule can be used as a template to construct another.

From DNA to DNA is called replication.

From DNA to RNA is called transcription.

From RNA to protein is called translation.

Today the central dogma has been extended.

There are examples where RNA templates RNA. And retroviruses can copy their RNA genomes into DNA by a mechanism called reverse transcription.
Making new molecules is called synthesis.

Certain proteins are required for the synthesis of both DNA and RNA.

DNA to DNA:

• A double helix is separated into two single strands.
• The single strands are used to template new double strands.
• Two identical DNA molecules are made.
DNA to RNA:

- A double helix is separated in a region by breaking the hydrogen bonds forming the base pairs.
- One strand of the DNA is used to template a single strand of RNA by moving along the DNA.
- These broken bonds form again at the end so that the double stranded DNA remains the same as before.
- A single strand of RNA is made.
The Genetic Code

As soon as Watson and Crick proposed the double helix model of DNA in 1953, scientists began to study the problem of how linear (or helical) DNA molecule could encode a linear protein molecule.

Crick’s approach was to assume that code reads blocks of letters. These blocks cannot be less than 3 letters long: $4$ and $4^2$ are both less than 20, whereas $4^3 = 64$ exceeds 20, since there are 20 amino acids.

The genetic code can be read from a single strand of RNA, and it is read from 5’ to 3’.
The code is a triplet code (codons): non-overlapping successive blocks of three letters are translated into amino acids.

There is a defined start or reading frame, start codon: AUG.

There are three triplets that cause protein translation to cease: UAA, UAG, and UGA.

With 64 possible words and 21 possible meanings, there is clearly the potential of different codons coding for identical amino acids.
mRNA

RNA that is translated into protein is known as messenger RNA, mRNA.

A shift of one letter in reading the same nucleic acid sequence results in a very different amino acid sequence.

The phase of codon reading is called the reading frame.

There are three reading frames going 5’ to 3’.

Reading the complementary DNA strand, there are another three read frames in the opposite direction.

Therefore there are a total of six possible reading frames for double stranded DNA.
Transfer RNA and Protein sequences.

mRNA is read to make protein. Amino acids are made available in the cell (some are synthesized by the cell itself).

With the amino acids and mRNA in the cell, how is a protein made?

Part of the answer lies with the so-called adapter molecule, another RNA molecule known as transfer RNA, tRNA.

Amino acids are linked to these smaller tRNA molecules of about 80 bases, and the tRNA then interacts with the codon of the mRNA.

Each tRNA has an anticodon of three bases which can base pair with a codon in mRNA. In this way, tRNA carries the appropriate amino acids to the mRNA.

At ribosome, mRNA is read and tRNA is utilized to make protein sequence.

Protein sequence then folds into folded protein.
Genes are not Simple

Starting and Stopping

As mentioned earlier, there are stop codons. There is also a so-called start codon, AUG, which codes for Met.

As is often in biology, the story is not so simple, and the details are highly dependent on the organism.

A molecular complex of several proteins called RNA polymerase is required to transcribe mRNA from DNA.
For efficiency and control, there are signals in DNA to start and stop RNA transcription.

The start pattern has specific small sequences in the DNA.

The RNA polymerase binds to these small sequences and then is in position to proceed down the DNA, transcribing it into RNA.

The small sequences the polymerase binds to are called promoter sequences.
Control of Gene Expression

Proteins from different genes exist in widely varying amounts: sometimes in ratio of 1/1000.

Gene expression could be controlled at two points: DNA → RNA or RNA → protein.

One common way to regulate a gene is by a repressor molecule (another protein), which affects the step DNA → RNA.
Suppose that the gene exists to process a molecule such as the sugar lactose.

When lactose is absent, a repressor molecule binds to the DNA, stopping the DNA → RNA step.

When lactose is present, it binds to the repressor and prevents it from binding DNA. When the expressed gene (the protein) has processed all the lactose, the repressor is no longer inhibited by lactose and the repressor again binds DNA, shutting down the transcription of the gene.

This clever scheme allows the organism to only make protein when needed, thereby saving much unneeded RNA and protein.
Split Genes

The genes for *prokaryotes*, organisms without a nucleus, are continuous subsequence.

However, the genes for *eukaryotes*, organisms with a nucleus, are interrupted by non-coding DNA that somehow disappeared in the mRNA.

The so-called *excons* become an uninterrupted sequence, whereas the so-called *introns* are spliced out and discarded.

Why did they evolve? How can we recognize genes in uninterrupted DNA? What are the signals for splicing out introns? These interesting questions do not yet have simple answers.
In viruses, where it is important to be compact, most of the DNA encoded genes. In higher organisms, this is far from the case. Humans have around 5% of the genome used in protein coding.

The function of much of the remaining DNA is unknown. Many people feel that much of it is Junk DNA not used for anything. Others think that this DNA has important biological functions that are not yet understood.
Jumping genes

In both prokaryotic and eukaryotic genomes, there are segments of sequences that move from place to place in the genome.

They carry required genes for their movement or transposition, hence the name jumping genes.

Since they propagate themselves, they create identical or similar segments of DNA in various places in the genome.

The role of these transposable elements is not very clear. Some have suggested that they are selfish DNA and exist only for their own well being.

For all the organisms examined, from bacteria to humans, we all have jumping genes.