Computational Biology
Lecture 1: Introduction
Saad Mneimneh

Quick Introduction (at the speed of light)

An important question in biological sciences in what makes life? A simple answer is Proteins and Nucleic Acids. Proteins are responsible for the body functions. Nucleic acids encode information necessary to produce proteins and pass this “recipe” to subsequent generations.

Most substances in our body are proteins. We have structural proteins responsible for building tissues, enzymes acting as catalysts for chemical reactions, and others such as transport proteins (hemoglobin) and antibody defense.

A protein is a long chain of Amino Acids. There are 20 different amino acids. Two amino acids bond together by a peptide bond. This is why an protein is called a poly-peptide chain.

But how to we get our protein? The answer lies in the Nucleic Acids. We have two kinds of nucleic acids: ribonucleic acids RNAs and dioxy-ribonucleic acids DNAs.

A DNA is also a chain of simpler molecules, namely sugar molecules. Each sugar molecule is attached to a base and this is what makes it different from the other sugar molecules. We have 4 bases, A, G, C, and T, thus the DNA can be viewed as a long sequence of 4 letters.

The DNA is actually a double stranded helix (discovered in 1953). The two strands hold together because each base in one strand bonds with a complementary base in the other. $A \leftrightarrow T$ and $C \leftrightarrow G$.

The RNA is similar to the DNA but it is single stranded, and every occurrence of a T is replaced by a U. U also bonds with A.

The genome is a collection of long DNAs that are called chromosomes. A gene is a stretch along a chromosome that encode the information for producing a specific protein. Only certain stretches on the chromosomes are genes. It is believed that 90% of the our DNA is *junk*. How does a gene produce a protein? By the process of *transcription*, the gene is made into an RNA. Then each 3 letters on the RNA, called *codon*, encode for an amino acid. Recall that there are only 20 amino acids and a codon can represent $4^3 = 64$ values. So many encodings lead to the same amino acid. This process is called *translation*.

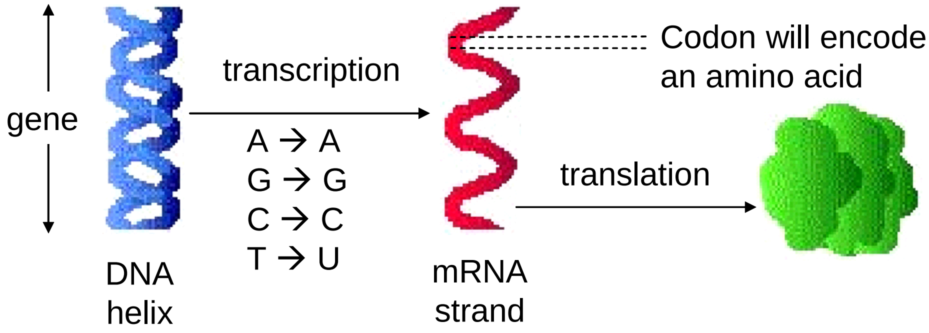

Figure 1: The Central Dogma of Biology

Some organisms, like bacteria, have a single chromosome which is a circular DNA molecule. Others, like humans, have many chromosomes, and each chromosome comes in two copies. The two copies are called homologous chromosomes. Although genes are the basic unit of heredity, the process of heredity occurs at the chromosome level. Mendel was among the first to study heredity and he came up with two laws:

- Each organism has two copies of a gene (one from each parent) on homologous chromosomes, and in turn, will contribute, with equal chance, only one of these two copies.
- Genes are inherited independently (not very accurate as we will see later).

Through a process called recombination, the two homologous chromosomes of the organism form a single one by taking only one copy of each gene from one of the two. This happens by a crossover mechanism where starting from one chromosome, a crossover to the other occurs, and so on... The obtained chromosome is passed to the child. Thus the child receives two homologous chromosomes (one from each parent) resulting in two copies of a gene one from each parent.

Why are genes so important? Because they dictate which proteins are produced. Moreover, most of the diseases have a genetic origin and are caused by the presence/absence of certain genes (for instance a mutation in a gene could trigger the disease). Therefore, the best cure for such diseases lies in finding the defective genes. But first, we have to find the genes.

Genetic Mapping

Specific genes, or specific gene mutations, are frequently the cause of diseases. In order to help us identify when a disease is associated with the presence of a particular gene, it is useful to have a map of the genome.

Genetic mapping is the process of determining the relative location of genes in a particular genome. By mapping the relative location of several genes, we can develop a map of the entire genome of a particular species e.g. humans.

Genetic mapping relies on externally observable genetic traits, or phenotypes, of a particular genome. By selectively breeding specific phenotypes together and examining the phenotypes of several generations of children, we can determine the frequency of the occurrence of particular phenotypes. After a large number of observations, we have a reasonable approximation of the probability of the occurrence of those particular phenotypes. Based on these probabilities, we can get an estimate of the distance between specific genes. As an example, let's take a genome that has n genes on a single chromosome and that recombination occurs randomly at only one point on the chromosome. Therefore, if the father's chromosome is $f_1 ... f_n$ and the mother's chromosome is $m_1 ... m_n$, a child can have either an $f_1 ... f_i m_{i+1} ... m_n$ or an $m_1 ... m_i f_{i+1} ... f_n$ chromosome, for some recombination position $0 \leq i \leq n$. As a result, every pair of parents can have $2(n+1)$ possible kinds of children (some might be identical) based on this single recombination position model. The probability of this recombination occurring at a particular position is $p = \frac{1}{n+1}$. The probability of two genes being separated by recombination is the probability that the recombination occurs in any position between them. This probability is expressed as $p = \frac{d}{n+1}$, where n is the number of genes in the genome and d is the distance between the two genes. Note that closer genes will have less chance of recombination (this is where the second law of Mendel is wrong, genes are not inherited independently if they are on the same chromosome). By starting with two different pure breed parents, say black and blue, we can concentrate on the states of two genes and observe the frequency of the states being different in the children (i.e. the frequency of recombination between those two genes). This will help us estimate p and therefore d. If we are able to determine the distance between all pairs of genes in our example genome, then we can use these distances to determine the exact sequence of the genes. However, real life is not as simple as our example.

In general, recombination occurs at an arbitrary number of positions in the genome. There are not always distinct phenotypes for a particular gene, or gene combination. This means that single changes cannot always be observed and in fact usually it takes multiple changes before there are any changes in phenotype. Moreover, some differences in the genotype are not reflected in any phenotype. It is highly probable that separate genes are distant and randomly distributed among multiple chromosomes. This makes our job even more difficult. On the other hand, if genes are particularly close, the order of the genes can not be determined because the two genes almost never occur separately.

References

Pevzner P., Computational Molecular Biology, Chapter 1.

Setubal J., Meidanis, J., Introduction to Molecular Biology, Chapter 1.