1 Biology/Chemistry Preliminaries

Here is a simplified view of the biological and chemical world we want to study. **Warning**: Some of the concepts and facts have been oversimplified (and, perhaps distorted), and almost every stated fact has exceptions. However, this should be adequate for us to get started.

- The main actors in the chemistry of life are the same in both simple and complex organisms; they are molecules called **proteins** and **nucleic acids**. Proteins govern all biological mechanisms, while nucleic acids encode the information necessary to produce the proteins. There are 2 kinds of nucleic acids – DNA and RNA.
- The (identical) genetic information in each cell of an organism is organized into a set of **chromosomes**; the complete set of chromosomes is called its **genome**.
- Each chromosome consists of a long polymer called **DNA**.
- Each chromosome has stretches of DNA called **Genes**, which are the basic units of inheritance [Mendel, 1865].
- DNA is a large double-stranded polymer and is made up of small molecules called **nucleotides** or bases.
- The bases are one of Adenine (A), Cytosine (C), Guanine (G), and Thymine(T). Thus, from a computational viewpoint, a DNA sequence can be thought of as a string on a 4-letter alphabet, and is completely specified by the base sequence along with an **orientation**. DNA sequences are oriented from the $5'$ end to the $3'$ end. The second strand is a **complementary** strand:

$$\begin{array}{cccccccccc} 5' & \mathrm{A} & \mathrm{C} & \mathrm{C} & \mathrm{T} & \mathrm{G} & \mathrm{A} & \mathrm{C} & 3' \\ & | & | & | & | & | & | & | & \\ 3' & \mathrm{T} & \mathrm{G} & \mathrm{G} & \mathrm{A} & \mathrm{C} & \mathrm{T} & \mathrm{G} & 5' \end{array}$$

- DNA has a **double helix** 3-dimensional structure [Watson and Crick, 1953].
- Every gene carries the code for a protein. The process of constructing a protein from the DNA sequence for a gene is the fundamental step that governs all biological mechanisms. The information flow in organisms is summarized by the **central dogma** [Crick, 1958]. While this has been extended in recent years, a simplified version is shown below:

$$\textbf{DNA} \rightarrow \textbf{RNA} \rightarrow \textbf{Protein}$$

- The process of transferring genetic information involves processes such as **Replication**, **Transcription**, and **Translation**.
- RNA is a single-stranded polymer made up of the following bases: Adenine (A), Cytosine (C), Guanine (G), and Uracil (U).
- A protein is made up of simpler molecules called **amino acids**. There are 20 possible amino acids. Thus, from a computational viewpoint, a protein can be thought of as a string of **residues** on a 20-letter alphabet $\{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\}$. Note that the alphabet is missing the letters B, J, O, U, X, Z. As with DNA sequences, protein sequences are also oriented; they go from the N-terminal to the C-terminal.
- Triplets of bases in genes called **codons** code for individual amino acids.
- The process of **Transcription** helps to decode the DNA sequence in a gene to make a type of **RNA** called the **messenger RNA** (mRNA).
- The process of **Translation** helps to make a **protein** from the mRNA sequence. The **Genetic Code** is used to decode the mRNA sequence. For example,

$$\text{mRNA} \quad \underbrace{\mathrm{UUU}}_{\text{Phe}}\underbrace{\mathrm{UAC}}_{\text{Tyr}}\underbrace{\mathrm{UGC}}_{\text{Cys}}\underbrace{\mathrm{GGC}}_{\text{Gly}}\mathrm{C}\cdots$$

- The presence of genes are indicated by **Start** and **Stop** Codons.

- A **reading frame** is one of the three possible ways of grouping bases on a DNA sequence to form codons in a DNA or RNA sequence. An **open reading frame** (ORF) in a DNA sequence is a contiguous stretch of bases staring at the start codon and consisting of an integral number of codons none of which are stop codons.
- Another type of RNA called the **transfer RNA** (tRNA) is the so-called adapter molecule used to assemble the specific amino acids that constitute a protein. This activity takes place on the **Ribosome**.
- The stretch of DNA that constitutes a gene may not be contiguous; the coding regions or **exons** are interspersed with non-coding regions or **introns**. In the process of making the mRNA from the DNA, the introns are **spliced** out.
- The linear sequence of residues in a protein is called its **primary structure**. Proteins actually fold in 3-dimensions and is important in imparting its functionality. A protein's **secondary structure** is formed through interactions between its residues and results in "local" regular structures such as α-helices and β-strands. **Tertiary structures** are formed by packing such structural elements into one or more compact globular units called **Domains**. The final protein may contain several polypeptide chains arranged in a **quaternary structure**. The folded protein froms functional regions called **Active sites**.
- A **plasmid** is a piece of (circular) extrachromosomal DNA, typically found in bacteria and often used as a cloning vector.
- A palindromic string reads the same backwards. The words "Malayalam" or "bib" are examples of palindromic strings. A **palindromic** DNA sequence is one that is equal to its reverse complement. "AGCT" is an example of a palindromic DNA sequence because when it is read backwards and complemented it reads the same.
- **Restriction Enzymes** cut DNA at palindromic **restriction sites**. A **Restriction Map** maps out all the restrictions sites in a given DNA sequence.
- **Sequencing** is the process of obtaining the base-pair sequence of a piece of DNA.
- Locating stretches of DNA sequences (such as a gene) on a chromosome is called **mapping**.
- **Sequence analysis** is performed by studying the statistical content of sequences. **Sequence Comparison** of two or more sequences is done by performing alignments of sequences. **Local Alignment, Global Alignment**, and **Semi-global alignments** are different kinds of alignments of interest.
- **Evolutionary Trees** are used to study the evolutionary relationships between species organisms. Such methods can also be extended to construct **Phylogenetic Trees** to study relationships between biological sequences.
- Since RNA is single-stranded, it can fold forming base-pairs with complementary nucleotides in another part of the sequence. The **RNA Secondary structure** describes these pairings and is of interest. This is considered as a step towards solving the much harder problem of determining the 3-dimensional structure.
- Three-dimensional structure of proteins are of tremendous significance. This is called the **protein folding problem**. Proteins form three basic kinds of secondary structures: α-**helices**, β-**sheets**, and **loops**.
- Proteins form higher-level structures called **motifs** and **domains**. A motif can be defined as a combination of a few secondary structures folded in a specific way and performing a specific function.
- The size or length of a DNA sequence can be measured by a process known as **gel electrophoresis**, which uses the fact that DNA is negatively charged and will move in an electric field and that the amount by which it moves depends on the size of the fragment. The gel method works well with small fragments.