Chapter 2

Biological Background

This chapter briefly discusses some basic and relevant biological concepts and terms required for proper understanding of our work [APS06, MDPM08, RJLS10, GSS91].

2.1 Basic concepts of a cell

Cell is the basic structural and functional unit of all living organisms. It may be regarded as a basic unit of biological activity. Each cell consists of jelly like material called protoplasm surrounded by a cell membrane. The protoplasm further has two components: the nucleus that contains the genetic material and the cytoplasm containing various other cell organelles like mitochondria, ribosomes etc. Cells of primitive organisms (such as bacteria) which do not have a nucleus are called prokaryotic cells and those of higher organisms which have a well defined nucleus are called eukaryotic cells.

There are four types of basic molecules present in a cell: sugars, lipids, amino acids and nucleotides. Sugar molecules besides being a source of energy for the cell, also play a structural role by forming the cell wall of the plant cell. Lipids are important constituent of the membrane of the cell and other cell organelles. Aminoacids are the structural units of proteins which are responsible for most of the biological processes in a cell. There
are twenty naturally occurring amino acids from which all proteins are composed. The linkage of one amino acid with another is through a peptide bond. Proteins, ‘also called polypeptide chains’ are long chains of amino acids. Each protein has a unique sequence of amino acids which determines its functionality. Proteins perform a variety of structural and dynamic functions in a cell.

2.2 DNA and RNA

The genetic information in a cell is stored in a long sequence of nucleotides \(^1\) called the Deoxy-ribose Nucleic Acid (DNA). DNA plays a key role in the transfer of genetic information from parent to its offspring.

\[\begin{array}{c}
\text{A} \leftrightarrow \text{G} \\
\text{T} \leftrightarrow \text{C}
\end{array}\]

Figure 2.1: Deoxyribose Nucleic Acid

It is made up of two strands wound together in a double helical structure. Each strand consists of sequences of nucleotides with bases adenine (A), thymine (T), cytosine (C) and guanine (G). These bases are repeated throughout the DNA strand millions of times. The bases present in one DNA strand forms hydrogen bonds with the bases in the other strand in a specific manner as shown in the Figure 2.1. ‘A’ pairs with ‘T’ only and ‘G’ pairs with ‘C’ only. Once the sequence of bases in one DNA strand is known, the sequence of the other strand can be constructed because of the specific base pairing. The two strands of DNA are thus said to be complementary and this property of DNA enables it to transmit genetic information. Further, each strand has a polarity, from head called

\(^1\) Nucleotides are molecules consisting of a nitrogenous base, a sugar group and a phosphate ion and are responsible for storage of information about an organism’s inherited characteristics.
the 5’ end and a tail called the 3’ end. The two strands are anti parallel (i.e. one strand runs from the 5’ to 3’ direction and the other from 3’ to 5’ direction). DNA of eukaryotes contain noncoding sequences called introns (or intervening sequences) that separate coding sequences called exons. Combinations of exons and introns form a gene which determine the sequence of aminoacids on a protein as shown in Figure 2.2. The part of DNA that codes for the formation of proteins is called an open reading frame (ORF). When read from the head region (5’ end) to the tail region (3’ end) the portion of DNA before the ORF is called the upstream region and the portion of DNA that comes after the ORF is called the downstream region of the DNA.

RNA or ribose nucleic acid is a molecule which is chemically similar to DNA. It plays a key role in the synthesis of various proteins in a cell. In some lower organisms it also acts as the carrier of genetic material. Unlike DNA, RNA is a single strand structure, constructed from a DNA strand having base Uracil (U) instead of the base Thymine. There are several types of RNA molecules. The RNA that codes for proteins is called messenger RNA (mRNA). Transfer RNA (tRNA) helps in aligning the amino acids according to the sequence present in the mRNA for the formation of the protein. Ribosomal RNA (rRNA) on the other hand is a major structural component of the protein synthesizing cell organelle called ribosome.

### 2.3 Gene and Gene expression

Gene refers to a functional unit of DNA. It is a sequence of bases that encodes a protein or an RNA molecule. Protein coding genes carry information for making proteins which determine various characteristics like colour of eye, hair etc. of an organism. The non protein coding genes code for RNA molecules. The physical characteristics of an organism i.e. what that organism looks like is called its phenotype and the genetic encoding of its phenotype is called its genotype. Set of all the genes in an individual is known as its
**genome.** The size of a genome may vary from 6000 genes in yeast to about 40,000 genes in human beings. **Gene Expression** is the process by which information from a gene (its sequence) is manifested into structure and functions of a cell. We say that a gene is expressed when the protein it codes for, is synthesized. Gene expression may also be referred to as the step in which the genotype of an organism is manifested into its phenotype i.e. the genetic information stored in the gene is expressed in the form of proteins that are responsible for the phenotype of the organism. Different subset of genes may be responsible for different phenotype of an individual. For example, a subset of genes responsible for the color of the eye may be different from the genes responsible for the height of an individual. Similarly in an apple tree, genes responsible for the shape of the fruit may be different from the ones that control its taste. Consider a gene that controls the height of an individual. The extent of expression of this gene determines the height. If the expression is within a normal range, the individual has a normal acceptable height. An over expressed gene might lead to a giant and an under expressed one leads to a dwarf.

The genotype of an organism interacts with the environment which thus influences the phenotype. In other words, the characteristics of an organism may be the result of the coordinated expression of one or several genes and their interactions with the environment. A gene may be highly expressed under certain conditions and may be suppressed under some other set of conditions. For example, an apple plant with the genotype of red juicy apple may give good quality apples in favorable conditions like significant amount of sunlight, air and water whereas the same plant under unfavorable conditions will yield bad quality apples.

Each cell of an organism contains the same set of genes. However different sets of genes are expressed in different cells. Also, within the same cell gene expression may vary with time and may also be affected by the internal and external state of the cell at that time. For instance, it is due to the coordinated action of several genes that the
The process of **gene expression** is a two step process that consists of transcription followed by translation. See Figure 2.3. **Transcription** is the process of transfer of genetic information from a portion of DNA into an mRNA molecule. **Translation** is the process of transfer of information from RNA to protein. **Transcription** is the first step in the formation of proteins from genes. It is carried out by an enzyme called RNA Polymerase. The process is initiated when one (or more) special proteins called **Transcription Factor (TF)** bind to one (or more) specific sequence(s) of nucleotides called the **Transcription Factor Binding Site(s) (TFBS)** on the promoter region of a gene. The enzyme RNA Polymerase moves along the strand of the DNA. As it encounters each DNA nucleotide, it adds the corresponding complementary RNA nucleotide to a growing mRNA strand. Once the stop signal is reached the newly constructed mRNA strand is released. Finally, it leaves the nucleus and serves as a template for the synthesis of protein in the cytoplasm at the ribosome.

**Translation** is the second step in the expression of genes. It involves reading the mRNA, conversion of the message carried in mRNA into amino acids and the synthesis of the corresponding proteins at the ribosomes. The genetic information based on the language of four bases (A, U, C, G) of mRNA is translated to a language of proteins consisting of amino acids. In other words, the sequence of nucleotides in the mRNA determine the sequence of amino acids in the synthesized protein. Each amino acid cor-
Figure 2.3: Two phases of Gene Expression

responds to a triplet of three nucleotide bases called a **codon**. 4 nucleotides can form 64 ($4^3$) possible codons out of which three (TAA, TAG and TGA) indicate the end of a protein sequence and are called the stop codons. All others code for a particular aminoacid. Most of the aminoacids are encoded by more than one codon. The codon AUG represents methionine and is also the translational **start signal**.

Translation begins when a tRNA (transfer RNA) molecule encounters the start codon on the mRNA. The tRNA moves up the sequence of the mRNA reading three nucleotides (codon) at a time. Each codon specifies the amino acid to be added to the protein sequence. The translation continues until a stop codon is encountered. The amino acid
chain is then released. This amino acid chain is nothing but the protein the gene codes for.

![Diagram of gene expression regulation](image)

**Figure 2.4: Effect of Conditions on Gene Expression**

Since the expression of a gene is controlled and regulated by one or more TFs and their binding to the TFBS in the promoter regions of the gene, it is expected that the genes that are coexpressed are regulated by the same set of TFs and hence have common TFBSs. In other words genes having similar expression profiles, thus belonging to the same bicluster, are considered to have a common regulatory mechanism or signature or motif in their promoter region [HZGD05].

Binding of Transcription factors with the TFBS may be regulated by many conditions as shown in Figure 2.4. Infact expression of one gene may be governed by the expression of another gene. Some genes may code for a protein which in turn may act as a transcription factor and regulate the expression of some other genes. The entire network is quite complex. Also, the gene to protein correspondence is not one to one. There are genes that may code for more than one protein. Ideally measurement of gene expression should be done by measuring the amount of protein produced. However, it is often easier to measure one of the intermediate product like mRNA to infer the gene’s expression level.
2.4 Microarray experiments and Expression matrix

A microarray experiment allows us to determine the expression levels of thousands of genes by measuring the amount of mRNA bound to each site of the microarray. Data resulting from a microarray experiment is represented as a gene expression matrix. Microarray experiments are based on the principle of hybridization. Linking of a DNA strand with an mRNA strand by hydrogen bonds is called hybridization. More the number of complementary bases present between the DNA and the mRNA strand, stronger is the hybridization. A microarray chip is a glass or a silicon slide [CQB04] which consists of an array of spots, where each spot contains multiple copies of a gene sequence known as probe. Thus in one experiment with a microarray chip with $n$ spots, we can study the expression levels of $n$ genes (probes) simultaneously. There may be thousands of such spots on a slide each containing millions of identical DNA molecules. From a sample of interest, e.g. a tumor biopsy, mRNA molecules (known as target) extracted and labeled with fluorescent dye are washed over the entire slide and is allowed to hybridize to the complementary gene sequences on the array. The sample contains more concentration of
the mRNA molecules corresponding to the genes highly expressed under the tumorous condition as compared to other genes. Thus the degree of hybridization will be more for such genes. By measuring the intensities of the fluorescent light emitted from each spot of hybridization, one can measure the amount of mRNA present in the sample. The images of the array are taken and analysed using image analysis software resulting in an intensity matrix.

Each intensity matrix is converted into a vector which corresponds to a column of the gene expression matrix with \( n \) rows or genes. Repeating the experiment for \( m \) conditions results in a \( n \times m \) gene expression matrix as shown in Figure 2.5. Each row in the matrix corresponds to the expression profile of a gene and each column corresponds to a sample or a condition [CST00]. The \((ij)^{th}\) entry of the expression matrix represents the expression of \(i^{th}\) gene under \(j^{th}\) sample.