Chapter 2

Background

2.1 Molecular Biology

The domains of biochemistry in general and molecular biology in particular, are concerned with the basic molecular principles of life. Biological objects interact with each other making possible all different forms of life. A central interest of molecular biology is the flow of information within an organism. Living organisms store all information that is necessary for growth, reproduction, and evolution in so-called genes on the DNA (sometimes RNA in simpler organisms). Deoxyribose nucleic acid (DNA) or Ribose nucleic acid (RNA) encodes the genetic instructions used in the development and functioning of all known living organism and thus act as informational molecules. The genetic information is encoded as a sequence of nucleotides: A (adenine), C (cytosine), T (thymine) G (Guanine) and U (Uracil). T appears only in DNA and U appears in RNA molecule. Watson-Crick model describes DNA molecules as double-stranded helices structure. It consists of two long polymer chains of nucleotide molecules attached with alternating sugars (deoxyribose) and phosphate backbone, which are both winded around a common axis that gives a double-helical structure to DNA. Each nucleotide molecule from one chain always bonds with a complementary nucleotide molecule from other chain and form a nucleotide pair called base-pair (bp). A base-pair is simply an interaction between the bases standing opposite of each other. These interactions are
based on hydrogen bonds. Erwin Chargaff suggested base pairing rule. According to the rule, A binds only to T (A-T) and C binds only to G (C-G), to form a base-pairs (see Figure 2.1, taken from\(^*\)). A DNA molecule of 100 bp thus consists of two antiparallel sequences, each 100 nucleotides (bases) long. The human genome roughly consists of \(1.3 \times 10^9\) of such base pairs\(^6\).

In contrast to DNA molecule, RNA is a single-stranded chain consists of four nucleotide molecules A, T, G and U. In living organism, there are four different types of RNA is available. Ribosomal RNAs (rRNA) are structural components of multi-protein complexes called ribosomes and protein synthesis takes place at the ribosomes. Messenger RNAs (mRNA) acts as carrier of genetic information from the genes to the ribosomes and transfer RNAs (tRNA) play the role of a translator, that translates the genetic information of the mRNA into a sequence of special bio-molecules called amino acids. Amino acids are the basic building blocks of protein. Protein molecules are polymers, i.e. consist of thousands or millions of atoms and responsible for all major cellular activities. It act as enzyme,

\(^*\)http://wps.prenhall.com/wps/media/objects/3513/3993159/bb2511.html
anti-bodies, regulatory substances, stabilisers, or carriers of other substances.

Genes are nothing but regions of the DNA and act as a repository of biological information which is necessary to build and maintain an organism's cells. It includes construction and regulation of proteins as well as other molecules that ultimately determine the growth and functioning of the living organism and transfer genetic traits to next generation. This is termed as central dogma of molecular biology. Entire DNA sequence of an organism do not play active role in cellular activities. In case of human genome, only 2-3% of the whole human DNA are functional. The functional part or the coding part of DNA is only responsible for protein synthesis. The remaining DNA does not encode for any protein. This DNA is sometimes referred to as “junk-DNA” or Non-coding DNA. Recent research reveals that junk-DNA plays critical roles in controlling how cells, organs and other tissues behave. Genes are templates for protein construction within a cell. Protein synthesis takes place within the cell through the process of transcription and translation. In transcription phase, a molecular complex called RNA polymerase-II creates a copy of a gene from the DNA to messenger RNA (mRNA) inside the nucleus. The mRNA travels from nucleus to the cytoplasm for protein synthesis, where it then binds with ribosome. Ribosome is a complex molecule based on ribosomal RNA (rRNA) and proteins. At the ribosome, mRNA is used as a blueprint for the production of a protein; this process is called translation. The mRNA moves along the protein synthesis site i.e. ribosomes, with a set of three-nucleotides called codons. Transfer RNA (tRNA) provides a compatible anticodon and is hybridised onto the mRNA. Finally, the amino acids bound to the RNA form polypeptide chain. This process continues until the translation process reaches a stop codon, which terminates the polypeptide synthesis. The entire process is called gene expression. A schematic drawing of the process of protein synthesis is illustrated in Figure 2.2 taken from National Health Museum®.

http://www.accessexcellence.org
2.2 Overview of Microarray Technology

Traditional experimentation system in molecular biology are capable of studying only a few genes in a single experiment. However, for a traditional methods, it is difficult to capture the dynamic behaviour or the activities of a gene that is going on inside a cell. DNA microarray technology provides a convenient and effective platform for monitoring activity of thousands of genes simultaneously. DNA microarray analysis is a fast and versatile approach to perform high throughput explorations of genome structure, gene expression, and gene function at both cellular and organism levels. Microarray analysis is a complex multi-step process involving various areas of expertise such as molecular biology, image analysis, computing and statistics.

There are five major steps in performing a typical microarray experiment\(^a\). The steps are illustrated in Figure 2.3 taken from\(^b\).

1. **Preparation of microarray**: In the preparation process, polymerase chain reaction (PCR) technique is used to amplify the DNA of interest using a universal primer or gene specific primers to generate thousands to millions of copies of a particular DNA sequence. The purity of the DNA fragments are then checked by sequencing or using an agarose gel through the estimation of the DNA concentration. The next step is spotting the DNA solution onto special glass slides coated with chemical materials such as polyethyleneimine polymer p-aminophenyl trimethoxysilane. Precision in spotting is achieved using precisely controlled robotic pins or other equivalent technology such as inkjet printing. The last step of manufacturing glass DNA microarrays is the post-print processing step involving drying of the DNA on the slide overnight at room temperature and the use of UV cross-linking to prevent subsequent binding of the DNA, and to decrease the background signal upon hybridisation of a labelled target.

\(^a\)http://grf.lshtm.ac.uk/microarrayoverview.htm

\(^b\)http://www.uni-koeln.de/med-fak/biochemie/transcriptomics/07_analysis.shtml
2. **Sample preparation and labelling:** Sample preparation begins with isolation of mRNA copies that represents genes, i.e., coding genes that expressed during sample collection. It is a very vital step as because the overall success of any microarray experiment highly depends on the quality of the RNA collected. Purity in terms of homogeneity or uniformity of the mRNA is an important factor for proper hybridization process, particularly when fluorescence is used, as cellular proteins, lipids, and carbohydrates can mediate significant nonspecific binding of labeled cDNAs to matrix surfaces. The mRNA extracted from both the target and the reference samples are then converted into complementary DNA (cDNA) using a reverse-transcriptase enzyme. To initiate cDNA synthesis, this step also requires a short primer. Next, each cDNA (target and reference) is labelled with a fluorescent cyanine dye (i.e. either Cy3 or Cy5).

3. **Hybridisation:** Hybridisation is the step of combining two complementary single stranded-DNA to form a double-stranded molecule. The labelled cDNAs (target and reference) are purified to remove contaminants such as
primers, unincorporated nucleotides, cellular proteins, lipids, and carbohydrates. After purification, the labelled cDNA is hybridised against cDNA molecules spotted already on a glass slide. Each molecule in the labelled cDNA will only bind to its appropriate complementary target sequence on the static array. Before hybridisation, the microarray slides are incubated at a high temperature with solutions of saline-sodium buffer (SSC), Sodium Dodecyl Sulfate (SDS) and bovine serum albumin (BSA) to reduce the background due to nonspecific binding.

4. **Washing:** To remove any unhybridized labelled cDNA from the array and to increase stringency of the experiment by reducing cross hybridisation, the slides are washed after hybridisation. The latter is achieved by either increasing the temperature or lowering the ionic strength of the buffers.

5. **Image acquisition and Data analysis:** The final step of microarray experiments involve image acquisition and data analysis of the array. The slide is dried at first and then scanned using a laser scanner to determine how much labelled cDNA (probe) is bound to each target spot. Laser excitation of the incorporated targets yields an emission with characteristic spectra, which is measured using a confocal laser microscope. Software used for microarray analysis often represents green spots as up-regulation a gene compared to control, red sport as down-regulation a gene in the experimental sample, and yellow to represent equal abundance in both experimental and control samples. In the data analysis phase, the relative expression levels of the genes in the sample and in the controlled populations can be estimated from the fluorescence intensities and colour for each spot. Based on the amount of probe hybridized to each target spot, information is gained about the specific mRNA composition and the representative in the sample. The logarithm of the ratio of raw red/green fluorescence intensities are taken to convert them into log intensities. In the case of microarray experiments, there are many sources of systematic variation that affect measurements of gene expression.
levels. The process of eliminating such variations allows appropriate comparison of data obtained from the two samples by using various normalization processes. The processed data, after normalization, can then be represented in the form of a matrix, often called gene expression matrix.

2.3 Gene Expression Data

Microarray is an indispensable technology in molecular biology that helps in assessing expression of a large number of genes under multiple conditions such as time-series, tissue samples (e.g., normal versus cancerous tissues), and experimental conditions. With the help of microarray experiments one can monitor simultaneously, the expression levels of several genes at a genome scale. To gain better understanding of a gene and its behaviour inside cell, various patterns can be derived by analysing the change in expression of the genes. An expression profile (of a gene or a sample) can be represented in vector space. For example, an expression profile of a gene can be considered a vector in \( n \) dimensional space (where \( n \) is the number of conditions), and an expression profile of a sample with \( m \) genes can be considered a vector in \( m \) dimensional space (where \( m \) is the number of genes). In the example given below, the gene expression matrix \( X \) with \( m \) genes across \( n \) conditions is an \( m \times n \) matrix, where the expression values for gene \( i \) in condition \( j \) is denoted as \( x_{ij} \):

\[
X = \begin{bmatrix}
  x_{1,1} & x_{1,2} & \cdots & x_{1,n} \\
  x_{2,1} & x_{2,2} & \cdots & x_{2,n} \\
  \vdots & \vdots & \ddots & \vdots \\
  x_{m,1} & x_{m,2} & \cdots & x_{m,n}
\end{bmatrix}
\]

Formally, it can be defined as:

**Definition 2.3.1 (Gene Expression Data)**: Let \( G = \{G_1, G_2, \ldots, G_m\} \) be a
set of \( m \) genes and \( R = \{T_1, T_2, \cdots, T_n\} \) be the set of \( n \) conditions or time points of a microarray dataset. The gene expression dataset \( X \) can be represented as an \( m \times n \) matrix, i.e., \( X_{m \times n} \) where each entry \( x_{i,j} \) in the matrix corresponds to the logarithm of the relative abundance of mRNA of a gene.

The expression profile of a gene \( i \) can be represented as a row vector:

\[
G_i = \begin{bmatrix} x_{i,1} & x_{i,2} & x_{i,3} & \cdots & x_{i,n} \end{bmatrix}
\]

The expression profile of a sample \( j \) can be represented as a column vector:

\[
G_j = \begin{bmatrix}
x_{1,j} \\
x_{2,j} \\
x_{3,j} \\
\vdots \\
x_{m,j}
\end{bmatrix}
\]

A subset of real gene expression data from a Homo-sapiens microarray dataset is given in Table 2.1.

### 2.4 Patterns in Gene Expression Data

Microarray data is essentially the logarithm of the ratio of raw red/green fluorescence intensities at a certain spot and is continuous in nature. The notion of pattern in microarray data introduced in\(^8\) as follows:

**Definition 2.4.1 (Expression Pattern)**: Given a gene \( G_i \), its expression values under a series of varying conditions or under a single condition form a range of real values. Suppose this range is \([a, b]\) and an interval \([c, d]\) is contained in \([a, b]\). Thus \( G_i \) is a vector of real numbers within the range \([a, b]\), denoted as \( G_i \subseteq [a, b] \), is called an item, meaning the values of \( G_i \) are limited inclusively between \( a \) and \( b \).

A set containing one single item is called a *pattern*. A set of several items,
Table 2.1: Sample gene expression data from Homo sapiens

<table>
<thead>
<tr>
<th>ORF</th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>C4</th>
</tr>
</thead>
<tbody>
<tr>
<td>GALNT5</td>
<td>-3.474</td>
<td>-3.837</td>
<td>-4.644</td>
<td>-5.059</td>
</tr>
<tr>
<td>APOE</td>
<td>-2</td>
<td>-1.943</td>
<td>-1.786</td>
<td>-1.737</td>
</tr>
<tr>
<td>IDH3B</td>
<td>1.449</td>
<td>1.299</td>
<td>0.993</td>
<td>0.832</td>
</tr>
</tbody>
</table>

Figure 2.4: Profile plot of Homo sapiens expression data

which come from different genes is also called a pattern. So, a pattern looks like:

\[
\{G_{i1}@[a_{i1}, b_{i1}], \ldots, G_{uk}@[a_{uk}, b_{uk}]\}
\]

where \(i_t \neq i_s, 1 \leq t, s \leq k\), if \(k > 1\).

Example patterns from a Homo sapiens microarray data (Table 2.1) and its corresponding profile plots are shown in the Figure 2.4.

It has been observed that from a biological point of view, patterns play an important role in discovering functions of genes, disease targets or gene interactions. A number of different patterns have been identified in biologically significant gene groups.

2.4.1 Shifting and Scaling patterns

In shifting patterns\(^8\) the gene profiles show similar trends, but distance-wise, they may be away from each other (see Figure 2.5). In terms of expression values, gene patterns follow an additive distance between them. Formally, shifting pattern can be defined as follows.

Definition 2.4.2 (Shifting Pattern): Given two gene expression profile \(G_i = \{E_{i1}, E_{i2}, \ldots, E_{ik}\}\) and \(G_j = \{E_{j1}, E_{j2}, \ldots, E_{jk}\}\) with \(k\) expression values, a profile
is called as shifted pattern, if expression value of $E_{i,k}$ can be related with $E_{j,k}$ with constant additive factor $\pi_k$ under $k^{th}$ condition. This can be written as follows.

$$E_{i,k} = E_{j,k} + \pi_k, \text{ for } i = 1 \text{ to } k \qquad (2.1)$$

Similarly, scaling patterns in gene expression follow roughly a multiplicative distance between the patterns. Scaling pattern can be defined as:

**Definition 2.4.3 (Scaling Pattern)**: Given two gene expression profile $G_i = \{E_{i1}, E_{i2}, \cdots, E_{ik}\}$ and $G_j = \{E_{j1}, E_{j2}, \cdots, E_{jk}\}$ with $k$ expression values, a profile is called as scaling pattern, if expression value of $E_{i,k}$ can be related with $E_{j,k}$ with constant multiplicative factor $\pi_k$ under $k^{th}$ condition. This can be written as follows.

$$E_{i,k} = E_{j,k} \times \pi_k, \text{ for } i = 1 \text{ to } k \qquad (2.2)$$

As shown in Figure 2.5, values of $G_2$ are roughly three times larger than those of $G_3$, and values of $G_1$ are roughly three times larger than those of $G_2$. In nature, it may happen that due to different environmental stimuli or conditions, the pattern $G_3$ responds to these conditions similarly, although $G_1$ is more responsive or more sensitive to the stimuli than the other two.
2.4.2 Coherent patterns

A group of genes showing similar pattern tendency across different conditions is called coherent. Such a group shows some kind of a co-expression in the expression profile. Co-expressed genes are likely to be involved in the same cellular processes. In practice, co-expressed genes may belong to the same or similar functional categories indicating co-regulated families\(^4\). Coherent gene expression patterns may characterize important cellular processes and may provide a foundation for understanding the regulation mechanism in the cells\(^9\). The patterns shown in Figure 2.5 are the examples of coherent patterns.

2.4.3 Co-regulated patterns

Often, coherent patterns are divided into two categories namely, positively regulated patterns and negatively regulated or inverted patterns. Sometimes, a group of genes that are positively or negatively regulated also called co-regulated genes. In Figure 2.4 genes GLANT5 and IDH3B show similar pattern or positively regulated patterns. On the other hand IDH3B or GLANT5 showing inverted or negative patterns with APOE. Biologically all three genes are very significant.

Thus, gene expression data analysis involves pattern finding. Data mining is the study of techniques that extract patterns from large amount of data. As a result, data mining provides the major tools for gene expression data analysis. Below we present a brief discussion of data mining techniques.

2.5 Data Mining

Data mining is a computational technique to analyze large volumes of data for finding relationship within the data that helps in predicting new fact such as how components of the data are related to one another. Fayyad, Piatetsky-Shapiro and Smyth in 1996\(^10\) defined data mining as: "The non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data". American statistician David Hand in 1998\(^11\) also defined data mining as: “A new
discipline lying at the interface of statistics, database technology, pattern recognition, and machine learning, and concerned with secondary analysis of large data bases in order to find previously unsuspected relationships, which are of interest of value to their owners". Data mining is an intermediate step in the KDD (Knowledge Discovery in Databases) process\textsuperscript{12} that consists of applying data analysis and discovery algorithms that produce a particular enumeration of patterns (or models) in the data. In general, the knowledge discovery process consists of an iteration sequence of the following steps:

- **Data cleaning**: handles noisy, erroneous, missing or irrelevant data.

- **Data integration**: where multiple, heterogeneous data source may be integrated into one.

- **Data selection**: where data relevant to the analysis task are retrieved from databases.

- **Data transformation**: where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operation.

- **Pattern evaluation**: identifies the truly interesting patterns representing knowledge based on some measures of interestingness.

- **Knowledge presentation**: where visualization and knowledge representation techniques are used to present the mined knowledge to the user.

### 2.5.1 Application of data mining

Today data mining offers value across a broad spectrum of industries.

1. **Marketing**: In marketing, the primary application\textsuperscript{13} of data mining is to analyze customer databases to identify potential customer groups and forecast their behaviour. Another popular application is market-basket data analysis systems, which extracts interesting patterns such as, "If customer bought X, he/she is also likely to buy Y and Z".
2. **Telecommunications**: Another application of data mining is the telecommunication alarm-sequence analyzer system (TASA). It was built in collaboration with a telecommunications equipment manufacturer and three telephone networks. Speciality of the system is that, it uses a novel framework for locating frequently occurring alarm episodes from the alarm stream and presenting them as rules. Large sets of discovered rules can be explored with flexible information-retrieval tools supporting interactivity and iteration. In this way, TASA offers pruning, grouping, and ordering tools to refine the results of a basic brute-force search for rules.

3. **Medical Application**: Medical applications are another fruitful area. Data mining can be used to predict the effectiveness of surgical procedures, medical tests or medications. Recently, it has also been used in Medical imaging applications to detect or predict diseases like cancer, which are sometime impossible for the human specialist to detect.

4. **Bioinformatics**: Data mining is used in the fields of biology and bioinformatics. Currently, data mining is extensively used in the analysis of gene expression data.

5. **Pharmaceutical**: Pharmaceutical firms are mining large databases of chemical compounds and genetic material to discover substances that might be candidates for development as agents for the treatments of diseases.

6. **Network Security**: Data mining is successfully used to predict the usage patterns in network to detect intrusion in the networks.

### 2.5.2 Data mining tasks

In general, data mining tasks can be classified into two categories.

**Descriptive mining**: It is the process of discovering the essential characteristics or general properties of the data in the database. Clustering, association and sequence mining are some of the descriptive mining techniques.
Predictive mining: This is the process of inferring patterns from data to make predictions. Classification, regression and deviation detection are predictive mining techniques.

There are several widely used data mining techniques. Traditionally, these techniques are used independently. These techniques include: classification, clustering, association rule mining, prediction and time-series analysis.

Classification: Classification\textsuperscript{12} is a supervised technique that partitions a given dataset into disjoint classes using a class attribute. A classifier model is built based on training data and later the model is used for predicting class of an unknown sample. The goal of classification is to analyze the training set and to develop an accurate description or model for each class using the attributes presented in the data. Many classifications models have been developed including neural networks, genetic models, and decision trees.

Clustering: Clustering\textsuperscript{12} is an unsupervised technique to group data into clusters with high intra-cluster similarity and low inter-cluster similarity. A similarity or distance measure is important criteria in deciding the quality of the cluster. To a large extent, quality depends on the appropriateness of the similarity measure for the data set or the domain of application. For example, clustering can be used to divide a population into distinct groups, such that each group can be treated with a different strategy. A number of clustering techniques are available. Partitioning methods, hierarchical methods, density-based methods, grid-based methods, and model based methods are some of the well known clustering techniques. The basic difference between classification and clustering is that classification assumes prior knowledge on class labels, while clustering does not assume any knowledge of classes.

Association Rules: Association rule mining\textsuperscript{12} is a data mining technique use to find interesting associations among a large set of data items. Association rule mining started with an initial idea to apply on market-basket analysis. In market-basket analysis, purchasing behaviour of customers are analyzed to
find association between different items that customers place in their "shopping baskets". The discovery of such association rules can help retailers in developing new marketing and placement strategies as well as logistics plan for inventory management that ultimately leads to business promotion. Association rules identify items that are frequently purchased together by customers. They make attempts to associate a product $A$ with another product $B$ so as to infer "whenever $A$ is bought, $B$ is also bought", with high confidence (i.e., the number of times $B$ occurs when $A$ occurs).

**Prediction:** Prediction techniques\textsuperscript{12} are based on some continuous valued attributes. The previous history of the attributes is used to build the model. This technique is commonly used for predicting product sales.

**Time-Series analysis:** Time-series analysis\textsuperscript{12} analyzes large sets of time series data to find regularities and interesting characteristics, including similar sequences or sub sequences, and sequential patterns, periodicities, trends and deviations. For example, one may predict trends in the stock values for a company based on its stock history, business situation, competitor performance and current market.

### 2.6 Discussion

The two classical data mining methods, i.e., data clustering and data classification, have been widely used to analyze gene expression data. These methods are valuable exploratory tools in data mining, and used successfully throughout the last two decades to explore biological knowledge from gene expression data. While classification helps in identifying genes responsible for diseases based on prior facts, clustering groups genes based on certain similarity measures into clusters that share common expression patterns. Unlike classification, clustering is effective in finding biologically significant groups of genes without any prior knowledge. These groups of genes involved in common functions and biological activities. However,
they are limited only to placing genes into disjoint groups that share certain characteristics. It has been observed that gene groups shares overlapping structures. Moreover, sometimes genes share similar expression patterns under a subset of given conditions. Biclustering\textsuperscript{20} is an extension of classical clustering, that has been successfully used in finding groups of genes having similar expressions under a subset of conditions.

Association rule mining\textsuperscript{21} is a relatively new and promising technique in the area of data mining and knowledge discovery. Association rule mining is a process that identifies links between sets of correlated objects in large datasets. Frequent itemset mining (we referred it simply as association mining in remaining of the thesis) is a sub-process of association rule mining technique, used to find relationship between the objects or items. Originally, the technique has been applied in market basket database and later extended to other application domains\textsuperscript{22,23,24} including neuroinformatics\textsuperscript{25}. However, not much work have been done so far to apply frequent mining or association mining in gene expression data analysis for finding gene regulatory networks or biclusters, with both positively and negatively regulated genes. Extension of classical association mining techniques for gene expression data analysis may suffer due to costly candidate generation phase and multi-pass nature of the techniques. Reducing the number of database passes and by removing the candidate generation phase may computationally improve gene expression data analysis based on association mining many folds.

In the next chapter, we present a new association mining technique called OPAM that needs only one pass over the database to generate all the frequent itemsets without any candidate generation.