Chapter 2

Gene Expression Pattern Identification

2.1 Microarray Technology: A Brief Overview

In 1970, Francis Crick introduced the central dogma of molecular biology [KW02] which has ever since, been one of the pillars of modern molecular biology. It pins down DNA (Deoxyribonucleic acid) as the carrier of genetic information and describes the unidirectional flow of information from DNA via RNA (Ribonucleic acid) to protein in three steps: Replication, Transcription, and Translation. This dogma is at the heart of bioinformatics which provides the framework to interrelate and interpret different types of data encountered in this field. The central dogma of molecular biology refers to the process of protein synthesis, which occurs in three major stages. The first stage, Replication, is the process which results in the duplication of the genetic information coded in DNA strands. The second stage, Transcription, is the transfer of information from double stranded DNA into single-stranded mRNA. The third stage, Translation, refers to the conversion inside the cell where mRNA is translated to produce a protein. Together Transcription and Translation constitute Gene Expression. Gene expression experiments provide a method to quantitatively measure the transcription phase of protein synthesis. The objective of gene expression experiments is the quantitative measurement of mRNA expression particularly under the influence of drug
or disease perturbations.

A DNA microarray or gene chip consists of an array of oligonucleotides or complementary DNA (cDNA) molecules of known composition chemically bonded to a solid surface (made of chemically coated glass, nylon membrane or silicon) [Ste06]. Gene chips are usually categorized into one of two classes, based on the DNA actually arrayed onto the support. An oligo array is comprised of synthesized oligonucleotides, whereas a cDNA array contains cloned or PCR-amplified cDNA (complementary DNA) molecules [KW02]. Both classes involve three common basic procedures [Ste06] which are depicted in Figure 2.1.

i Chip manufacture: A microarray is a small chip, onto which tens of thousands of DNA molecules (probes) are attached in fixed grids. Each grid cell relates to a DNA sequence. The DNA on the array are referred to as probes and the labeled DNA in solution as target.

ii Sample preparation, labeling, hybridization and washing. The first step is the extraction of RNA from the tissue of interest. Next, two mRNA samples are reverse-transcribed into cDNA and labeled using either fluorescent dyes (Cy3 and Cy5) or radioactive isotopes. It is then hybridized with the probes on the surface of the chip. Hybridization is the step in which the DNA probes on the glass and the labeled DNA (or RNA) target form heteroduplexes via Watson-Crick base pairing. After hybridization, the slides are washed (using a low-salt wash or with a high-temperature wash) to remove excess hybridization solution from the array. This ensures that only the labeled target on the array is the target that has specifically bound to the features on the array. This step also reduces cross-hybridization. This process is illustrated in Figure 2.2 reproduced from http://ttitan.biotech.uruc.edu/cs491jh/slides/cs491-le1.ppt.

iii Image Acquisition: In this step, an image of the surface of the hybridized array (chip) is produced by scanning the chip to read the signal intensity that is emitted from the fluorescent dye of the heteroduplexes on the array where the target has bound to the probe.

The image acquisition process is shown in detail in Figure 2.3 and has been repro-
Figure 2.1: Steps in a microarray experiment (Courtesy of [Ste06]). The Cy3 and Cy5 in the diagram refers to the mRNAs dyed using the two fluorescent dyes of Cy3 and Cy5.

After hybridization, the slides are scanned using a laser device to determine the amount of fluorescent label that is attached to each cDNA on the slide. The amount of fluorescence is displayed as a cell on a matrix corresponding to the spot on the original slide. The images output from the scanner are colored according to a standard where a higher level of fluorescent label (enhanced gene expression) is colored red, a lower level (repressed level of gene expression) is colored green and equal levels are yellow.

The digital image obtained from the image acquisition step is converted into numerical measures of hybridization intensity for each channel on each feature [Ste06]. The image is analyzed by (i) Gridding: Identify spots (this step can be automatic, semiautomatic or manual); (ii) Segmentation: Separate spots from background using fixed circle, adaptive circle, adaptive shape or histogram methods; (iii) Intensity extraction: Obtain mean or median of pixels in spots and (iv) Background correction: can be either local or global. The microarray data thus generated is then cleaned, transformed and normalized to resolve any errors, noise
Figure 2.2: Hybridization of probe array

Figure 2.3: The image acquisition process.
and bias introduced by the microarray experiments [Ste06]. The logarithm of the raw intensities are taken to convert them into log intensities. Once expression data is obtained from the microarray images using various standardization and normalization procedures [Ste06], the information embedded in the data has to be analyzed.

A gene is expressed in a cell when the protein it codes for is actually synthesized. About 10,000 genes are expressed in an average human cell. The set of (say 10,000) numbers that indicate the expression level of each of these genes is called the expression profile of the cell.

The power of a microarray experiment derives from the fact that many thousands of different DNA molecules are bonded to a single array. So it is possible to measure the expression of many thousands of genes simultaneously, conveniently and efficiently.

### 2.2 Gene Expression Data

Gene expression of a gene refers to effective production of the protein that a gene encodes. A microarray experiment assesses a large number of DNA sequences (genes) under multiple conditions such as time-series, tissue samples (e.g., normal versus cancerous tissues), and experimental conditions. A gene expression data set from a microarray experiment may be considered as a $G \times T$ matrix $D_G$ as shown in Equation 2.1, the rows of which represent expression patterns of a set of $G$ genes $\{g_1, \ldots, g_G\}$, and the columns represent expression profiles of a set of $T$ samples, $S = \{s_1, \ldots, s_T\}$ and each cell $g_{i,j}$ is the expression level of gene $g_i$ (where $1 \leq i \leq G$) on sample $s_j$ (where $1 \leq j \leq T$).

$$D_G = \begin{bmatrix} g_{11} & g_{12} & \cdots & g_{1T} \\ g_{21} & g_{22} & \cdots & g_{2T} \\ \vdots & \vdots & \ddots & \vdots \\ g_{G1} & g_{G2} & \cdots & g_{GT} \end{bmatrix} \quad (2.1)$$

Cluster analysis starts with this gene expression matrix and calculates proximity among genes. Clustering algorithms group genes which are similar based on a
proximity measure into the same cluster. Therefore, similar genes are grouped into the same cluster and dissimilar genes are grouped into different clusters.

During the last few years, several significant coherent pattern identification techniques have been developed under the categories of gene based, sample based and subspace clustering approaches. The next section is dedicated to reviewing the popular algorithms.

2.3 Gene Expression Pattern Identification using Data Mining

Cluster Analysis is an unsupervised grouping technique used to group similar objects (in this case genes) into disjoint sets based on their attribute (condition) similarities. Clustering identifies genes with similar expression profiles (co-expressed genes). Co-expressed genes have similar expression profiles, while a coherent expression pattern represents the common trend among expression levels for a group of co-expressed genes. Furthermore, co-expressed genes in the same cluster are likely to be involved in the same cellular processes, and a strong correlation among expression patterns of the genes indicates co-regulation. In practice, co-expressed genes may belong to the same or similar functional categories and indicate co-regulated families [AMS94]. Various gene clustering methods have been used to identify co-expressed genes and discover coherent expression patterns. A cluster of genes contains a group of co-expressed genes and the coherent expression pattern is obtained as the mean (the centroid) of the expression profiles of the genes in the cluster.

Gene clustering techniques are divided into three different types: gene-based clustering where the genes are treated as objects while the samples are features, sample-based clustering in which the samples can be partitioned into homogeneous groups where the genes are regarded as features and the samples as objects, and subspace clustering in which either genes or samples can be regarded as objects or features. Both gene-based and sample-based clustering approaches search exclusive and exhaustive partitions of objects that share the same feature space. Subspace clustering algorithms capture clusters formed by a subset of
genes across a subset of samples. Throughout the work reported in this thesis, we use the gene-based clustering approach. In gene-based clustering, similar rows (genes) are grouped together into unique clusters. The premise is that each cluster shows a similar temporal expression pattern as shown in Figure 2.4 and may represent a group of functionally related genes i.e., a biological module.

2.4 Proximity Measures

A microarray experiment compares genes from an organism under different development time points, conditions or treatments. For a $T$ condition experiment, a single gene has a $T$-dimensional observation vector known as its gene expression profile. A similarity (or dissimilarity) measure is a real-valued function that assigns a real number as a similarity or dissimilarity value between any two expression vectors. Therefore, to identify genes or samples that have similar expression profiles, appropriate similarity (or dissimilarity) measures are required. Some of the commonly used distance metrics are: Euclidean distance, Pearson's Correlation coefficient and Spearman's Rank-order Correlation Coefficient [Ste06]. The Euclidean distance measure imposes a fixed geometrical structure and finds clusters of that shape even if they are not present. It is scale variant and cannot detect negative correlation. Euclidean distance gives the distance between two genes but does not focus on the correlation between them. Pearson's Correlation, on the other hand, retains the correlation information between two genes as well.
as the regulation information. However, since it uses the mean values while computing the correlation between genes, a single outlier can aberrantly affect the result. Spearman's Rank Correlation is not affected by outliers, however there is information loss w.r.t. regulation since it works on ranked data.

### 2.5 Gene Expression Data Clustering Approaches

Data mining techniques have proven to be useful in understanding gene function, gene regulation, cellular processes and subtypes of cells. According to [Ste06], most data mining algorithms developed for gene expression time series deal with the problem of clustering. Clustering identifies subsets of genes that behave similarly along a course of time. Categorization of gene expression data clustering techniques is discussed next.

#### 2.5.1 Partitioning Approaches

k-means [McQ67] is a typical partition-based clustering algorithm which divides the data into pre-defined number of clusters in order to optimize a predefined criterion. The major advantages of it are its simplicity and speed, which allows it to run on large datasets. However, it may not yield the same result with each run of the algorithm. Often, it can be found incapable of handling outliers and is not suitable to detect clusters of arbitrary shapes. A Self Organizing Map (SOM) [Koh95] is more robust than k-means for clustering noisy data. It requires the number of clusters and the grid layout of the neuron map as user input. Specifying the number of clusters in advance is difficult in case of gene expression data. Moreover, partitioning approaches are restricted to data of lower dimensionality, with inherent well-separated clusters of high density. But, gene expression data sets may be high dimensional and often contain intersecting and embedded clusters. QT (quality threshold) clustering [HKY99] is an alternative method of partitioning data, invented for gene clustering. It requires more computing power than k-means, but does not require specifying the number of clusters apriori, and always returns the same result when run several times. The distance between a point and a group of points is computed using complete linkage, i.e., as the
maximum distance from the point to any member of the group [ESBB98]. A hierarchical structure can also be built based on SOM such as Self-Organizing Tree Algorithm (SOTA) [DC97]. Recently, several new algorithms such as [HVD01] and [THHK02] have been proposed based on the SOM algorithm. These algorithms can automatically determine the number of clusters and dynamically adapt the map structure to the distribution of data. Herrero et al. [HVD01] extend the SOM by a binary tree structure. At first, the tree only contains a root node connecting two neurons. After a training process similar to that of the SOM algorithm, the data set is segregated into two subsets. Then the neuron with less coherence is split into two new neurons. This process is repeated level by level, until all the neurons in the tree satisfy some coherence threshold. Other examples of SOM extensions are Fuzzy Adaptive Resonance Theory (Fuzzy ART) [THHK02] which provide some approaches to measure the coherence of a neuron (e.g., vigilance criterion). The output map is adjusted by splitting the existing neurons or adding new neurons into the map, until the coherence of each neuron in the map satisfies a user specified threshold.

### 2.5.2 Hierarchical Approaches

Hierarchical clustering generates a hierarchy of nested clusters. These algorithms are divided into agglomerative and divisive approaches. Unweighted Pair Group Method with Arithmetic Mean (UPGMA), presented in [ESBB98], adopts an agglomerative method to graphically represent the clustered dataset. However, it is not robust in the presence of noise. In [ABN+99], the genes are split through a divisive approach, called the Deterministic-Annealing Algorithm (DAA). The Divisive Correlation Clustering Algorithm (DCCA) [BD08] uses Pearson's Correlation as the similarity measure. All genes in a cluster have highest average correlation with genes in that cluster. Hierarchical clustering not only groups together genes with similar expression patterns but also provides a natural way to graphically represent the data set allowing a thorough inspection. However, a small change in the data set may greatly change the hierarchical dendrogram structure. Another drawback is its high computational complexity.
2.5.3 Density Based Approaches

Density based clustering identifies dense areas in the object space. Clusters are highly dense areas separated by sparsely dense areas. DBSCAN [EKSX96] was one of the pioneering density based algorithms used over spatial datasets. In [JPZ03a], Jiang et. al. propose the Density-Based Hierarchical clustering method (DHC) to identify co-expressed gene groups. It can identify embedded clusters in the dataset and can also handle outliers. It can effectively visualize the internal structure of the data set. A kernel density clustering method for gene expression profile analysis is reported in [SZCS03]. An alternative to this is to define the similarity of points in terms of their shared nearest neighbors. This idea was first introduced by Jarvis and Patrick [JP73]. A density-based approach discovers clusters of arbitrary shapes even in the presence of noise. However, density-based clustering techniques suffer from high computational complexity with increase in dimensionality (even if spatial index structure is used) and input parameter dependency.

2.5.4 Model Based Approaches

Model based approaches provide a statistical framework to model the cluster structure in gene expression data. The Expectation Maximization (EM) algorithm [DLR77] discovers good values for its parameters iteratively. It can handle various shapes of data, but can be very expensive since a large number of iterations may be required. In [TH09], a signal shape similarity method used to cluster genes using a Variational Bayes Expectation Maximization algorithm [BG03]. A model-based approach provides an estimated probability that a data object will belong to a particular cluster. Thus, a gene can have high correlation with two totally different clusters. However, the model-based approach assumes that the data set fits a specific distribution which is not always true.

2.5.5 Graph Theoretical Approaches

In graph-based clustering algorithms, graphs are built as combinations of objects, features or both, as nodes and edges, and partitioned by using graph theoretic
algorithms Graph theoretic algorithms are also used for the problem of clustering cDNAs based on their oligo-nucleotidc fingerprints [HSL+99] CLuster Identification via Connectivity Kernels (CLICK) [SS00] is suitable for subspace and high dimensional data clustering The Cluster Affinity Search Technique (CAST) by [BDSY99] takes as input pairwise similarities between genes and an affinity threshold It does not require a user-defined number of clusters and handles outliers efficiently But, it faces difficulty in determining a good threshold value In CAST, the size and number of clusters produced is directly affected by the fixed user-defined parameter t and hence, apriori knowledge of the data set is required To overcome this problem, E-CAST [BPC02] calculates the threshold value dynamically based on the similarity values of the objects that are yet to be clustered

2.5.6 Soft Computing Approaches

Fuzzy c-means [Bez81a] and Genetic Algorithms (GA) (such as [BMM07] and [MMB09]) have been used effectively in clustering gene expression data The Fuzzy c-means algorithm requires the number of clusters as an input parameter The GA based algorithms have been found to detect biologically relevant clusters but are dependent on proper tuning of the input parameters

The current information explosion, fuelled by the availability of the World Wide Web and the huge amount of microarray experiments being conducted, have led to ever-increasing volume of data Therefore, there is a need to introduce incremental clustering so that updates can be clustered in an incremental manner Though a lot of research has been performed on incremental clustering in other application domains, incremental clustering of gene expression data has not been exploited much yet

2.5.7 Incremental Algorithms

In [EKS+98], the authors present an incremental clustering approach based on the DBSCAN [EKSX96] algorithm A one pass clustering algorithm for relational datasets is proposed in [TS08] Rough set theory is employed in the incremental
approach for clustering interval datasets in [ANS06]. In [LLF+04b], an incremental genetic k-means algorithm is presented. In [RRAR06], an incremental gene selection algorithm using a wrapper-based method that reduces the search space complexity since it works on the ranking directly, is presented.

2.6 Discussion

From the discussion above, we conclude that various clustering algorithms require different types of input parameters and clustering results are highly dependent on the values of parameters. Gene expression data has coherent patterns embedded in the full gene space, identification of which is an important research field. Coherent genes may indicate co-regulation and hence fall under the same functional classification. Clustering algorithms that do not require the number of clusters as an input parameter and are robust to noise are of utmost importance. Clustering algorithms are sensitive to the proximity measure chosen. In this thesis, we present several clustering methods and all of them use a proximity measure developed by us which is introduced in the next chapter.