Chapter 1

1 Introduction

One of the leading research areas of recent times is Bioinformatics. It integrates diverse fields which includes computer science and informatics, biology, statistics, applied mathematics and artificial intelligence so as to provide solutions involving biological problems at the molecular level. With the help of data mining techniques and statistical methods, it has become possible to organize, analyse and interpret biological data with an aim to uncover and identify interesting patterns in the underlying data. The major areas of Bioinformatics concern microarray and gene expression data analysis, protein structure and primary genome sequence. The genome provides only static information whereas the gene expression data analysis produced from the microarray experiments provides dynamic information about cell function. The measurement of the activity (expression) of thousands of genes at once so as to create a global picture of cellular function is known as gene expression profiling, which is done with the help of microarrays. Analysis and interpretation of microarray gene expression data for the purpose of extracting biologically relevant knowledge are a fundamental task in data mining.

Two classical data mining methods: data clustering [1] and data classification [2] have been widely used to analyze gene expression data. Clustering is the process of grouping data objects into a set of disjoint groups, called clusters, so that objects within a group have high similarity to each other, while objects in separate groups are more dissimilar. Classification refers to a procedure that assigns data objects to a set of

classes. In this approach, prior knowledge about the objects is directly exploited by the algorithm.

The gene expression data in the cancer gene expression database (CGED) is derived from static expression experiments [3] that analyze samples from many individuals. Although different classification methods from statistical and machine learning area have been applied to cancer classification, these methods were not designed to handle high dimensional data efficiently and effectively with minimum computation time. What is required at present to address the fundamental problems relating to cancer diagnosis and drug discovery are high accuracy classification and clustering approaches that reveal biological information so as to produce effective cures for diseases. For this reason ensembles of classifier algorithms and clustering algorithms are preferred since it generates an overall improved partitioning of the dataset, being the consensus of the participating individual algorithms. More importantly, a user does not have to select a particular clustering algorithm and is saved from the risk of making a poor choice.

1.1 Cluster Analysis

Gene expression data are generated from high-throughput microarray technologies and they are often presented as matrices of expression levels of genes under different conditions. One of the major objectives of gene expression data analysis is to identify groups of genes having similar expression patterns over the full space or subspace of conditions. This may reveal natural structures and identify interesting patterns in the underlying data. A cluster of genes can then be defined as a set of biologically relevant genes which are similar based on a proximity measure. The goal of clustering is to partition the elements into subsets called clusters, so that two criteria are satisfied: *homogeneity* – elements in the same cluster are highly similar to each other; and *separation* – elements from different clusters have low similarity to each other.

Analysis of gene expression data can be done using supervised or unsupervised methods. In supervised methods, the dataset is partitioned into disjoint classes using a class attribute. A classifier model is built based on training data (data sample plus

correct class labels) and later the model is used for predicting the 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 present in the data. Many classifications models have been developed including neural networks, genetic models and decision trees.

In unsupervised methods, the dataset is partitioned into disjoint groups called clusters with high intra-cluster similarity and low inter-cluster similarity. A similarity or distance measure is an important criterion in deciding the quality of the cluster. To a large extent, quality depends on the appropriateness of the similarity measure used for the data set or the domain of application. A number of techniques are available for the purpose of clustering gene expression data. 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.

The supervised and unsupervised methods are not free from their own biasness and cannot provide an accurate analysis of high dimensional data. Therefore, we turn our focus on ensemble methods to improve the overall prediction accuracy by combining the output of several algorithms. Moreover, we can combine ensemble approaches to gain further improvement by building an ensemble of ensembles known as the meta-ensemble.

1.2 Gene Expression Clustering

The basic steps involved in applying a clustering algorithm for the purpose of clustering gene expression data can be summarized as follows:

1. *Feature selection*: This process identifies the most effective subset of the original features to be used in the clustering process, after filtering out the irrelevant and redundant genes.

2. *Clustering process*: This step involves the utilization of a suitable clustering algorithm for the underlying data distribution of the gene expression dataset so as to generate a good clustering of the data. A clustering algorithm uses a proximity measure and a search method to find the optimal or sub-optimal groupings in the dataset according to some clustering criterion. A proximity measure quantifies the similarity (or dissimilarity) of two data points and the clustering criterion is based on the working definition of a cluster and/or an expected distribution of underlying data in specific application domain.

3. *Cluster validation*: Cluster validation is the assessment of the clustering obtained in terms of quality, in order to get a correct informative biological explanation of the gene cluster, with the help of validation indices.

The issue of grouping genes having similar expression patterns is generally addressed by employing traditional clustering algorithms such as partitioning, hierarchical, density-based, model-based, graph theoretic and soft computing clustering algorithms.

1.3 Microarray Datasets – Gene, Cancer Data and Protein Interaction Data

Using a microarray, it is possible to examine the expression levels of thousands of genes across different developmental stages, clinical conditions or time points simultaneously. The real-valued gene expression data is obtained in the form of a matrix where the rows refer to the genes and the columns represent the conditions. 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.

DNA microarrays can be used to determine which genes are being expressed in a given cell type at a particular time and under particular conditions. This allows us to

compare the gene expression in two different cell types or tissue samples, so as to determine the more informative genes that are responsible for causing a specific disease or cancer. The accuracy of microarray dataset analysis depends on both the quality of the provided microarray data and the utilized analysis approach or objective. However, the problems and issues arising out of the curse of dimensionality, the small number of samples and the number of irrelevant genes make the classification task of a test sample more challenging. The objective of clustering expression profiles of tumours is to determine new disease (cancer) classification. Clustering aims at dividing the data points (genes or samples) into groups (clusters) using measures of similarity, such as correlation or Euclidean distance.

Proteins are the building blocks of all living organisms. The *Central Dogma of Biology* describes the formation of protein inside a living organism, as shown in *Figure* 1.1.

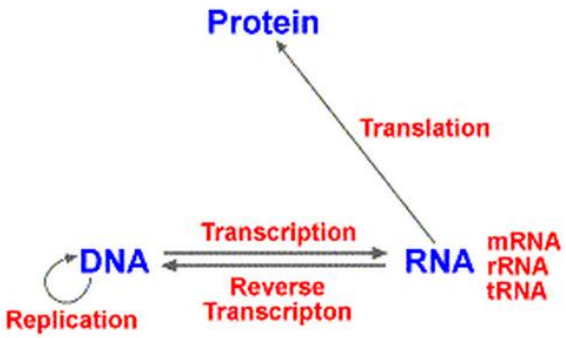

Figure 1.1: The Central Dogma

The double-stranded DNA molecule is partially unzipped and an enzyme called RNA polymerase copies the gene's nucleotides one by one into an RNA molecule, called the messenger RNA or mRNA. This process is called transcription. The mRNA is a small, single stranded sequence of nucleotides which moves out of the nucleus. Outside the nucleus, another set of proteins reads the sequence of mRNA and gathers free floating amino acids to fuse them into a chain. The sequence of the mRNA determines the order in which each amino acid is incorporated into the growing protein. The process of translating the mRNA sequence into a protein sequence is called translation.

Proteins are the cause of many diseases. If a newly discovered protein complex gets correctly classified, then the task becomes easy for the drug analyst to discover new drugs. Therefore, correct identification of protein complexes becomes a very challenging task as it guides the analysts to discover appropriate drugs. Protein interaction network data is often noisy and hampers the accurate detection of protein complexes. One way to compensate for this would be to combine multiple datasets for the purpose of identification of meaningful complexes using some clustering technique.

1.4 Motivation

Based on a comprehensive literature survey, the following conclusions have been arrived at:

- The effectiveness of a clustering technique is highly dependent on the proximity measure used by the technique. Choosing or finding an appropriate proximity measure is a challenging task.

- Most existing clustering techniques are either dependent on input parameter(s) or stopping criteria for discovery of the "true" number of clusters, which is a major task and hence a challenge.

- Gene expression data often contain clusters which are "highly connected" [4], intersected or even embedded [5]. Hence the clustering algorithm should be capable of effectively managing such a situation.

- There are two main issues involved in the design of ensembles: (i) the diversity of the algorithms to form a potentially accurate ensemble, and (ii) integration of the outputs of the clustering and classification algorithms to obtain a consensus.

- The diversity and the quality of the base classifiers and clustering algorithms highly influence the quality of results produced by the ensemble. Hence the

influence of diversity and quality of the base classifiers on high dimensional clustering needs to be examined.

- It is necessary to develop a better understanding of the relationship between the performance of various combination and consensus functions and the basic properties (diversity and quality) of ensembles.

- A combination rule that will work for any type of data (numerical, categorical and mixed) and will not be influenced by the dimensionality or biases of the participating classifiers or clustering methods is required.

- Development of a robust, integrated tool to construct gene-gene network (both expression as well as regulatory) using an ensemble approach that combines multiple sources of biological information. The tool should provide validation support using co-expression similarity, semantic similarity and sequence similarity.

- Designing a generic combination function to support cluster analysis using ensemble approach based on multiple sources of input biological data to extract non-trivial patterns of high biological significance, is another important research issue.

- Use of ensemble methods for the analysis of gene expression data in view of providing biologically relevant information, by incorporating unlabeled data, would be of help to biologists.

- An important research area in unsupervised ensemble approach is represented by stability-based methods for the assessment of the number and the reliability of the clusters discovered by clustering algorithms. There is a need to assess the reliability of the discovered clusters in bioinformatics problems, as well as the proper selection of the "natural" number of clusters in the underlying data.

- Identification of protein complexes of high biological significance from large number of Protein-Protein Interaction (PPI) datasets based on supervised and unsupervised ensembles is desirable.

- The noise in the PPI datasets hampers the detection of accurate protein complexes. One way to compensate for the lack of interaction coverage would be to combine multiple datasets to purify the dataset for the purpose of identification of meaningful complexes.

- A single protein may be involved in a number of functions and hence occurring in many complexes. It becomes imperative to assign importance to proteins based on the number of interactions in which they participate and then rank the protein complexes with an aim of isolating the protein for the purpose of producing effective cures for diseases.

An in-depth study of the above issues will provide useful guidance for applying ensemble techniques in practice.

1.5 Contributions

In this thesis a comprehensive literature survey on clustering algorithms is first carried out, followed by an exhaustive and comprehensive survey of the ensemble methods used in supervised, unsupervised and semi-supervised approaches. The survey discusses certain issues in ensemble learning and highlights the importance of ensemble diversity, ensemble framework, various ensemble combination methods and their applicability to clustering microarray and high dimensional gene expression data.

For the first task classification techniques using ensemble approach are applied in gene expression data analysis. An empirical study of various existing supervised classifiers and their ensembles is first performed to identify their pros and cons. The idea was to develop a cost effective ensemble method which is not influenced by the biasness of the base classifiers and should show improved detection rates consistently. This is achieved by combining the base classifiers from different classification families into an ensemble, based on a simple estimation of each classifier's class performance. An improvement in the classification accuracy is seen by using this approach and it is extended to develop a meta-ensemble by combining the results of

the proposed ensemble with the results of the best performing ensemble methods. Experimental results are presented to establish the effectiveness of the proposed model, validated over nine cancer datasets.

Analysis of cancer datasets demands high accuracy and hence it is essential to validate the results of clustering with the help of external and internal validation. A method is devised involving unsupervised ensemble approach to improve upon the analysis of the results that were obtained from the previous task on cancer datasets. To assist in the cluster validation process of cancer datasets, it is proposed to integrate existing biological knowledge such as the GO database. The clustering results arrived at are validated using external validity measures such as semantic and sequence similarity measures to ascertain whether the clusters obtained are biologically significant. Internal validity measures such as Dunn index, silhouette width and the homogeneity index are used to evaluate the visual separation of the clusters obtained from a clustering algorithm. Since the dataset is highly correlated, additional stability measures and biological validation measures in the form of biological homogeneity index and biological stability index are used to obtain biologically relevant clusters. The approach was tested on several benchmark cancer datasets and the experimental results have been found to be excellent.

Finally, an ensemble for protein complex identification is contributed since ensembles have been able to improve the robustness and stability of clusterings by combining the output of several algorithms. The errors made in clustering by one algorithm are likely to be averaged out by the correct clustering of another, so that the overall clustering accuracy is improved and a final unbiased decision can be arrived at. The ensemble is developed based on the limitations uncovered after evaluating and reviewing eight state-of-the-art techniques for computational detection of protein complexes. The PPI network data is taken as input by n consistently well-performing clustering algorithms which generate n individual complexes. Then the node with the highest degree is identified from the complex set C_1 and the corresponding protein complexes are identified based on this node. A common set of nodes (proteins) P_i, are identified from the corresponding complex sets with the condition that each member element of P_i is present in atleast two complexes and then complete its edges to generate a complex, which is arrived at by consensus among the corresponding complexes. For

the same common set of nodes P_i identified earlier, the edges are also completed based on GO similarity, to obtain a protein complex. Now, both the complexes so obtained are superimposed to obtain a final complex, using a simple logic to decide whether to retain or discard an edge between a set of nodes. This process of identifying the node of the highest degree from the next set of complexes and the process of identifying protein complexes through consensus and through the use of GO similarity is repeated for each corresponding complex. The end result of this phase is interacting protein complexes with overlapping proteins.

Finally, the validation phase evaluates protein complexes predicted by comparing them to a set of gold standard protein complexes and also checks for the biological relevance of the predicted protein complexes.

1.6 Organization of the Thesis

The thesis is organized as follows:

- *Chapter 2* gives the background of the study. It discusses microarray technology and presents a literature survey on supervised, unsupervised and semi-supervised ensemble approach on microarray gene expression data and its helpfulness in gene expression pattern identification using data mining techniques. A brief discussion on protein complex identification from protein-protein interaction data is also included.

- *Chapter 3* proposes a supervised ensemble technique referred to as SD-Enclass whose performance in the classification accuracy was better as compared to the single best model in the combination. A Meta-Ensemble was then developed using the combination method proposed which gave significantly better results than using Boosting, Bagging or Stacking alone.

- In *Chapter 4* presents an extension of the work done in *Chapter 3*. A method is devised involving unsupervised ensemble approach so as to improve the analysis of previously obtained cancer data results. The clustering results

arrived at are validated by integrating existing biological knowledge in the form of semantic and sequence similarity measures.

- *Chapter 5* presents an evaluation and review of eight state-of-the-art algorithms used for computational detection of protein complexes. Based on the issues and limitations uncovered, an ensemble approach is developed where Gene Ontology information is integrated with PPI network data so as to improve the protein complexes prediction accuracy.

- Finally, the concluding remarks are given in *Chapter 6*.