Chapter 1

Introduction

1.1 Introduction

A living being is said to possess characteristics such as growth, reproduction, metabolism and homeostasis through which it sustains life [140, 157, 164]. All organisms are made up of cells which are capable of growth and reproduction. These cells are made up of DNA, which stores genetic material. For a cell to perform any life sustaining activity, this genetic material becomes functional by means of processes described by the *Central dogma* [22]. The segments of DNA responsible for producing such materials are called genes and the functionally activated molecules are called proteins. Proteins and genes are responsible for the phenotypic traits within an individual in addition to performing life sustaining activities [21]. Figure 1.1 shows the relationship between genes, proteins and phenotypic traits in a living individual.

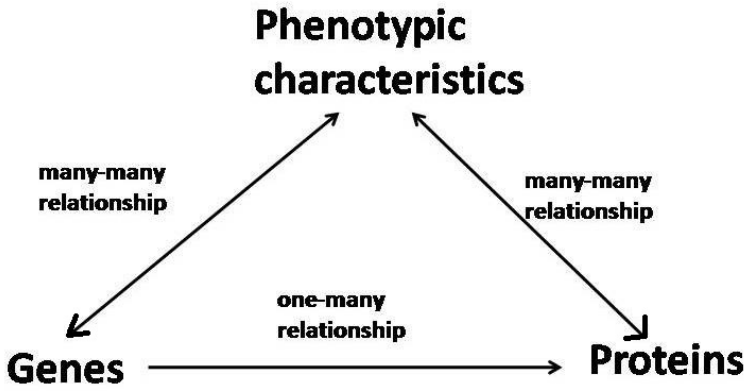

Figure 1.1. Relationship between genes, proteins and phenotypic traits.

1.1.1 Protein Protein Interaction Data Analysis

The proteins which are formed in the living cell during the process described by the Central Dogma have different combining capabilities. They have affinity only for certain other kinds of proteins. When two or more such proteins coordinate activities with each other, they exhibit interesting behaviors, which can be interpreted as the expression of phenotypic traits in the individual. A number of experimental techniques are available to detect such interactions. The most commonly used are the Yeast two-Hybrid (Y2H) and the Tandem Affinity Purification methods (TAP). The Y2H method is based on the color change reaction that takes place during the activation of transcription factors associated with the proteins of interest. The TAP method also works on the same principle by considering the reaction between a tagged protein of interest and its binding partners. The interaction between proteins are then made available to the researchers in the form of Protein Protein Interaction (PPI) data. Several repositories such as MIPS [96], DIP [122], HPRD [62] maintain such datasets and provide them upon request.

Computational researchers have moulded this data as graphs by representing them in network forms, technically referred to as Protein Protein Interaction Network (PPIN). The most popular form of analysis that can be carried out on such networks is the identification of complexes, which are nothing but groups of similar proteins acting together to accomplish certain functions. One can leverage this information to infer properties of unknown proteins or they can be used as aid to infer disease related information [15, 44].

1.1.2 Gene Expression Data Analysis

Analyzing diseases from protein interactions seem to be superficial and incomplete. This is because one cannot infer much from the availability of interacting proteins, and experimental techniques used for detecting such interactions are error-prone. Therefore, one needs to consider the genetic background to definitively establish any findings related to diseases. The genetic material within the DNA tends to become functional only if it has the capacity to produce certain functional or non-functional products. The process by which the information in genes governs the creation of meaningful products is known as gene expression. Each cell contains a multitude of genes. However, not all of them express themselves at the same time and place. These genes tend to form functional products only under the action of certain catalyst like material, technically referred to as transcription factors [76]. A number

of methods are available to identify the expression level of genes under different conditions. Historically, the most common has been the microarray technology [47], where the genes under consideration are placed at the intersection points of a grid like structure known as the microarray chip. A color dye solution is then used to dip the chip, thereby promoting the hybridization process. The expression levels of genes are then measured from the amount of color dye deposited in the chip. Another most recent technique that is being used to measure gene expression is the RNA-Seq technology [16], where mRNA molecules are first shortened into nucleotide base pairs, which are then aligned with known nucleotide base pairs to determine gene expression levels.

The expression levels of genes are determined under different time points emphasizing their time-dependent nature. Such experiments result in time dependent expression data [6]. In some cases, the expression levels of genes are recorded under different time points, under different conditions giving rise to three-dimensional gene expression datasets [172]. Gene Expression Omnibus-NCBI [9] and Array-Track [145] are the most common repositories to get access to gene expression data.

Analyzing the expression level of genes helps unravel a part of the mystery behind the working of the living body. Gene expression analysis has been useful in gene function prediction, identification of regulatory relationships among the genes, and in disease diagnosis [14].

1.1.3 Data mining

The amount of biological data is increasing at a very rapid rate. The enormous growth in the quantity and quality of biological data such as gene expressions, sequence data, molecules and pathway information provide for huge but possibly raw and noisy knowledge sources at the genetic level. The proper interpretation of such information can assist scientists in revealing the mysteries behind the activities of genes. In order to study the vast domain of biological data, one has to depend on specialized computer aided processing. The most relevant techniques which can be used here come from data mining because such techniques deal with automatic discovery of interesting patterns from huge amount of data [29]. Various approaches to data mining such as regression analysis, clustering, classification and association rule mining have been successfully used in inferencing quality information from raw biological data. Regression analysis has been used to estimate the dependency of specific variables on independent parameters [43]. Classification is the process

of learning from a set of labeled objects in a way that makes it possible to label any new unlabeled object in the future [43]. Clustering is a type of unsupervised learning that groups objects based on similarities between the groups they belong to [43]. Association rule mining deals with quantifying the relationship between elements that occur together, as if in a transaction [43].Data mining approaches have been primarily used in protein interaction and gene expression data anlaysis.

1.2 Issues and Challenges

Substantial research has been undertaken in the area of protein protein interaction data and gene expression data analysis. However, there are issues which have not been addressed properly during the years. The following are some of the issues that I have addressed during my PhD period.

- The lion's share, a researcher's attention in PPI data analysis usually goes to detection of groups of similar proteins or protein complexes. Most complex finding approaches proposed in the literature are proximity based and the results are biased towards the datasets under study. Hence, such methods show inconsistent performance when used on a range of datasets.

- Another issue which arises during performance analysis of many complex detection methods is that they usually apply to yeast datasets only. If the same methods are used on human PPI dataset, they show wide performance variations. Therefore, a method which can detect protein complexes reliably from both yeast as well as human PPI datasets would be more useful.

- Usually, a protein complex finding method works on the sequential principle of optimizing several objective functions. Very few methods have been proposed till date considering the parallel approach to objective function optimization. When two or more objective functions need to be simultaneously maximized (minimized) to find high quality complexes, a multi-objective parallel optimization technique would be more appropriate.

- The noisy nature of PPI data makes it less reliable for detection of quality complexes. One can integrate GO information with the PPI data to get quality complexes. A semantic similarity metric which gives the distance between any two nodes in the GO network can be used along with PPI data to enhance the reliablity of complex detection.

- Proteins and genes are known to be associated with phenotypic characteristics in a living being. The intermittent nature of expression of genes in cells gives rise to different features in the body. Therefore, analysis of their properties in a subspace of conditions can be used to derive meaningful inferences related to the phenotypic variations occurring in a person.

1.3 Research Objectives

State-of-the art research shows the utility of PPI and gene expression data analysis in predicting the behavior of genes during the progression of diseases. For better inferencing, the techniques available to analyze biological data, in this case PPI and gene expression data must be highly reliable, especially when used in the case of human datasets. With this objective in mind, I have worked my thesis towards accomplishing several sub-objectives, which would lead to more effective and extensive analysis of both PPI and gene expression data. The sub-objectives of my research are as follows.

- The first objective of my research was to study the performance of existing protein complex finding methods over a number of datasets and observe the variations in the performance obtained.
- After analyzing the performance limitations of the existing methods, I proposed a complex finding method which would work well on a range of datasets.
- My third objective was to propose a complex finding method that works well for both yeast and human PPI datasets. It was then extended to rank the obtained complexes in terms of their relevance w.r.t. a disease query.
- My next objective was to find interesting and novel disease genes which undergo changes from one stage of the disease to another stage. This work was as an extension of the module extraction technique. The module extraction technique was most appropriate here because higher the coherence between modules, the more interestingly such genes behave during the disease progression.
- The final objective of my thesis was to analyze the discontinuous nature of genes by proposing a subspace based gene module extraction technique and using it to discover the behavioral changes of biomarker genes from a healthy to a diseased person.

1.4 Research Contributions

To address a few of the issues arising in PPI and gene expression data analysis, I have proposed the following solutions during the course of my PhD work.

1.4.1 Protein complex extraction techniques that show consistent performance across a range of yeast datasets as well as human dataset.

In order to find quality complexes with minimal parameter tuning, I have proposed a method called CNCM. This method is based on the connectivity of nodes with that of the seed node in the PPI network and uses just two parameters. It has been tested on four yeast datasets and has been shown to perform well for all the datasets in terms of precision, recall and F-measure.

I have also proposed a method called DCRS which is based on the parallel evaluation of a set of objective functions. The method uses a combination of both topological and biological functions which need to be simultaneously maximized, and is based on the use of NSGA II optimization technique. The performance of my method has been tested for human dataset in terms of sensitivity, positive predictive value and accuracy. The contribution made in this work is listed under Publications section as Publication No.1 & 2.

1.4.2 Quality protein complex detection methods and disease associated complex inferencing

A few studies have been conducted to analyze the role of proteins in diseases. In this context, I have proposed a method called CSC based on seed expansion strategy which could identify quality complexes. The performance of CSC has been shown to be better in terms of sensitivity, positive predictive value and accuracy for the human PPI dataset. I have analyzed the obtained complex members w.r.t. Alzheimer's disease from the disease gene-central gene point of view.

In addition, I have proposed a semi-supervised method called ComFiR to further enhance the quality of complex detection. This method uses the computation of semantic similarity between proteins as a member function during complex finding. The accuarcy of complexes obtained using ComFiR is far better than all other existing methods till date in case of human PPI dataset. This work is then ex-

tended to rank complexes which may be relevant for a query disease. Publication No.4 & 5 in the Publication section lists down the publication contribution made in this work.

1.4.3 A gene module extraction technique to find interesting behavior of genes involved during disease progression

Any disease tends to progress in stages over time. The study of behavior of genes which change their activity patterns during the stages is very useful and informative. In this work, I have proposed a gene module extraction technique which is able to find highly similar genes for both non-metastasis and metastasis stages of breast cancer. The pathway details and other associated information of the biomarkers discovered from our work have been validated by literature research. S.No.6 of Publication list gives the contribution of this work.

1.4.4 A subspace based gene module extraction technique to study the progression of disease from normal to diseased state

The intermittent expression property of genes has been considered in this work. I proposed a gene module extraction technique based on a subset of acting conditions. This work relies on multi-edge connectivity among genes in the gene gene network and extends the idea to find coherent gene modules in both control and diseased states of Parkinson's disease. It is then extended to find correspondence between modules in both the stages and in the process, a few novel disease genes were identified. The behavior of these disease genes were then studied in terms of pathway and interacting partners, and also from expression point of view as the disease starts developing in a healthy person. The contribution of this work is given as S.No. 3 & 7 under Publication list.

1.5 Organization of the Thesis

Chapter 2 presents background of protein protein interaction data analysis, gene expression data analysis and the use of data mining in these analyses. It also discusses various datasets and tools that are used in my research work. Chapter 3

discusses two protein complex finding methods, the first one is named CNCM and the second one is named DCRS. The next chapter, i.e., Chapter 4 reports two more complex finding methods namely CSC and ComFiR. It also discusses the relation between complex finding methods and Alzheimer's Disease. Chapter 5 proposes a gene module extraction technique and its application in breast cancer. The next chapter, Chapter 6 discusses a subspace based gene module extraction technique and its application to Parkinson's Disease. The final chapter, Chapter 7 presents the concluding remarks for my PhD work and also highlights the future directions in my research work.