Chapter 2

Background

2.1 Introduction

The rapid growth of gene expression data has attracted the attention of researchers to study and analyze the roles of interactions among biomolecules, whether it be at gene level, protein level, metabolite level or a combination of these. Understanding the functions of these molecules requires deep insight. All these molecules arise as a result of a mechanism known as the *Central dogma*. These molecules, when they interact with one another, give rise to networks like protein protein interaction networks, gene gene networks or metabolite networks. The combination of all these interactions in the living cell consitute the "interactome". We now discuss the process by which these molecules are produced in the living body in detail.

2.1.1 Central Dogma

A cell is considered the smallest independent unit capable of life. Information from a cell is passed from one cell to another through a molecules called DNA(Deoxyribose Nucleic Acid). DNAs are polymers made up of nucleotides, comprising of a five carbon sugar, attached to a phosphate group and a base. The sugar attached to the phosphate group forms the skeleton and the base pairs such as Adenine (A), Thymine (T), Guanine (G) and Cytosine (C) determine the biological information being coded by the chemical structure of the DNA strand [1]. Physically, a DNA is a helical structure present inside the nucleus of a cell, comprising of coding and non-coding regions [2]. The coding regions of DNA are known as exons while the

[1] https://www.ncbi.nlm.nih.gov/books/NBK26821/
[2] https://www.ncbi.nlm.nih.gov/books/NBK21171/

non coding regions are known as introns. Both these segments are known as genes. However, the non-coding genes do not have the ability to produce proteins. Hence, the segments with coding capacity are frequently referred to as genes, which are actually responsible for all the biological activities occurring in the cell. A diagram representing the structure of DNA along with the introns and exons is shown in Figure 2.1 and 2.2.

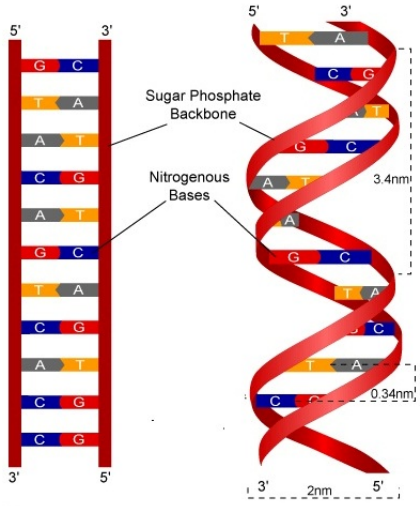

Figure 2.1. DNA structure (source-http://theinvestigation.yolasite.com/dna-structure.php, accessed on-14/12/2017)

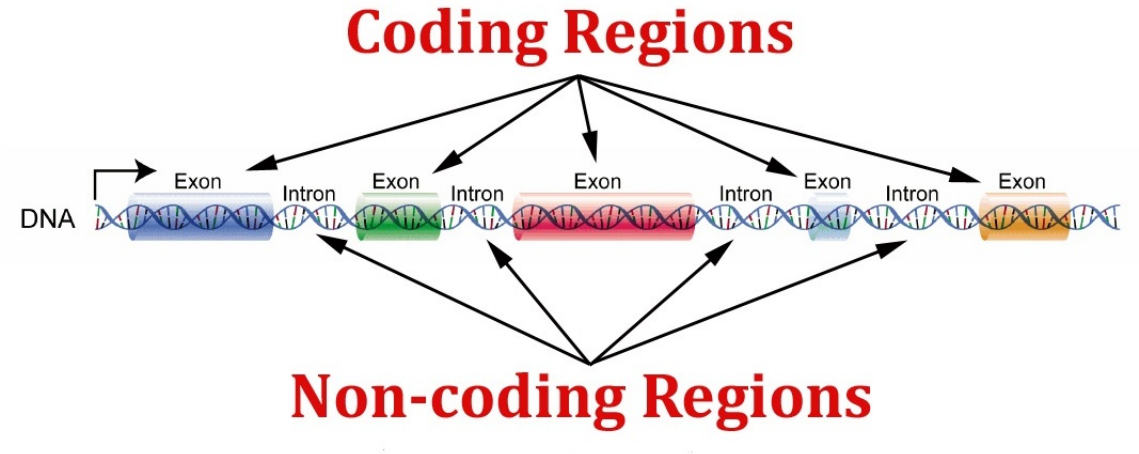

Figure 2.2. Coding and non-coding regions of DNA (source-https://scienceoveracuppa.com/tag/fragile-x-syndrome/, accessed on-14/12/2017)

A gene is said to be expressed if it is able to produce its corresponding protein. However, information from the DNA cannot be simply used for protein manufacture. It involves a series of steps. The first step is the transfer of information from

the double helix DNA to a single strand mRNA molecule by means of complementary base pairing assisted by the RNA Polymerase II enzyme. This step is known as transcription, where DNA molecule gets transcribed into RNA molecule. The next and final step involves reading the bases of mRNA molecule in triplets of code, known as codon. These codons correspond to 20 different amino acids. The triplets in mRNA serve as basis for linking a large number of amino acids giving rise to different proteins [18]. This step is known as translation, where nitrogeneous bases gets converted into proteins. The two steps together make up the *Central Dogma* of Molecular Biology and make the existence of life, and heredity from generation to generation possible. A diagrammatic representation of *Central dogma* is shown in Figure 2.3.

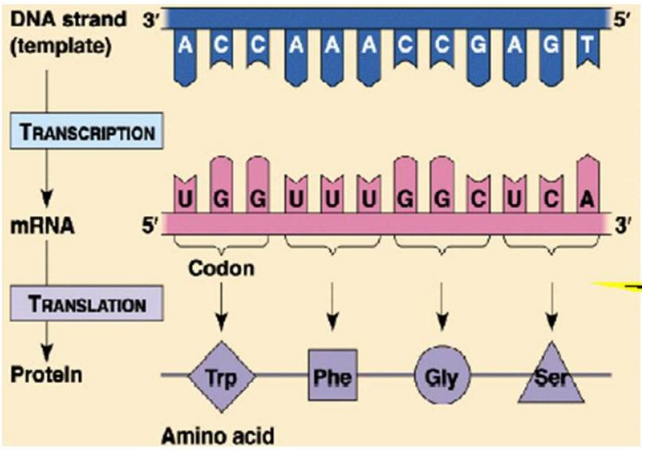

Figure 2.3. Central dogma-Protein synthesis from DNA (source-http://fourier.eng .hmc.edu/bioinformatics/intro/node8.html, accesesed on 14/12/2017)

2.1.2 Protein Protein Interaction Data

A combination of amino acids produces a multi-molecular compound known as protein. Proteins are involved in regulating various activities in the living body. However, they need to function in coordination with other proteins to function. These interactions among proteins are a result of biochemical events, which are regulated by certain electrostatic forces[3]. Interactions among proteins represent the most crucial process occurring in the living body. A study of these interactions help infer the functions of each protein in a group; it can also help predict the functionality of some uncharacterized proteins based on the nature of its interaction

[3]https://en.wikipedia.org/wiki/Protein-protein_interaction

with known proteins. Some of the effects of protein protein interaction, as given in [105], is listed here.

- Protein interactions, being dynamic in nature, have the ability to modify the rates as well as properties of certain binding substrates or enzymes. For example, succinate thiokinase interacts with α-ketoglutarate dehydrogenase to decrease the K_m for succinyl coenzyme.

- Interactions among proteins provide a platform for substrate chanelling. For example, tryptophan synthetase is brought about by a two-step process-formation of indole from indole 3-glycerol phosphate, which is then converted to tryptophan.

- These interactions can also lead to the formation of new binding sites. When interaction between α and β subunits of E. coli takes place, an ADP site gets formed at the interface.

- They can also lead to temporal inactivity in some enzymes. This is the case when trypsin tries to interact with trypsin inhibitor or when phage T7 gene 1.2 protein interacts with E. coli dGTP triphosphohydrolase.

2.1.2.1 PPI Detection Methods

A living cell consists of nearly 20000 genes which produce nearly 500,000 proteins. Around 80% of these proteins are known to act in coordination with other proteins. Studying the interactions among such a huge number of proteins is difficult. Certain *in vitro, in vivo* and *in silico* methods have been designed to detect protein interactions. A given procedure of PPI detection, if carried out in a contolled environment outside the living body is known as an *in vitro* technique. It is the complete opposite of an *in vivo* technique which is carried out inside a living organism. Some researchers have also come up with an *in silico* technique, where PPI is detected via computer simulation. We now discuss some PPI detection techniques.

- Tandem Affinity purification (TAP) : This method involves a double tagging stage on the considered protein. The protein of interest is first coated with a TAP tag which consists of calmodulin binding peptide (CBP) at the tobacco etch virus protease (TEV protease) cleavage site. This TAP tagged protein then binds to IgG site of some beads. This binding results in the breakage of TAP tag and the protein of interest binds reversibly to another IgG associated

molecule. Thereafter, the considered protein is washed off throughly from the affinity column and its binding partners are then monitored[4].

- Co-immunoprecipitation : In this process, an antigen is attached to a target protein involved in complex formation. This newly formed complex is then precipitated, and an antibody binding protein is brought in contact with it. The final step is to determine the identity of the antigen using electrophoresis.

- Protein-fragment complementation assays (PCA) : These are a class of assays used in studying PPIs in any living cell or in *in vitro* conditions [98]. Protein detection using mass spectrometry can be performed either by peptide fingerprinting or shotgun proteomics. In peptide fingerprinting, the washed complex is separated by SDS-PAGE. The gel is Coomassie stained and each protein is digested with the help of some enzyme and then analyzed using mass spectrometry. On the other hand, in shotgun proteomics, the whole washed complex is digested before analysis.

- Phage display : This is a comparatively new method, where the protein of interest and its partner are determined from DNA and gene levels [136]. This process is completed by a validation step using the yeast two-hybrid method.

- Yeast two-Hybrid : This is an *in vivo* method for complex detection and involves two domains-one is the DNA binding domain (DBD) which binds to the DNA in bait protein, and the other is the activation domain (AD), which activates the transcription process of DNA in prey protein [56]. The change in color of the prey protein indicates the presence of an interaction between the two proteins.

- Fluorescence resonance energy transfers (FRET) : It uses the correlation between time and each photon to predict PPIs [86].

- *In silico* methods : A number of *in silico* methods exist, which support experimentally detected PPIs. These include sequence [51], structure [88], gene fusion [28], chromosome proximity [162] and phylogenetic tree [123] based.

A summary of the PPI detection techniques is given in Table 2.1.

[4]https://en.wikipedia.org/wiki/Tandem_affinity_purification

Table 2.1 Summary of PPI detection techniques

Category	**Method**	**Working**	**Disadvantage(s)**
In vitro	TAP-MS	Works in two steps-purification process followed by mass spectroscopy	Tags may affect the protein expression levels
	Co-immunoprecipitation	Uses target specific antibodies to find proteins attaching to the protein of interest	Low affinity or transient proteins cannot be detected.
	PCA	Can be used for any weight protein to detect PPIs among them	Works well only with small samples
	Phage display	Incorporates protein and genetic materials in a single phage	Comparatively new, yet to be established
In vivo	Yeast two-hybrid	Analyzes a protein with a random set of potential partners	Large number of false positives and false negatives.
	FRET	Uses time related information for each protein	Time and concentration dependent.
In silico	Structure based, sequence based, gene fusion based, phylogenetic tree based	Evolution based and supports multi-domain functionallity of proteins	Requires a powerful system for large interactome

2.1.2.2 Types of PPI data

Protein protein interaction datasets which are published in the literature are mostly the outcomes of wet lab experiments. These data are mostly in two column format, representing the protein pair among which the interaction takes place. In this type of dataset, the confidence of each edge among the pairs is taken to be the same. This is the *unweighted PPI dataset*. There are also cases when the reliability score of each interaction is given in a third column. This type of dataset belongs to the *weighted* class. The third column in such a dataset corresponds to the confidence score with which the two proteins interact in nature. Higher this score, stronger is the interacting pair.

2.1.2.3 Protein protein interaction data analysis

The interactome of proteins in a living body can be represented in the form of a network called the PPI network (PPIN). This network consists of vertices corresponding to proteins and edges corresponding to interactions between them. An

example PPIN is shown in Figure 2.4.

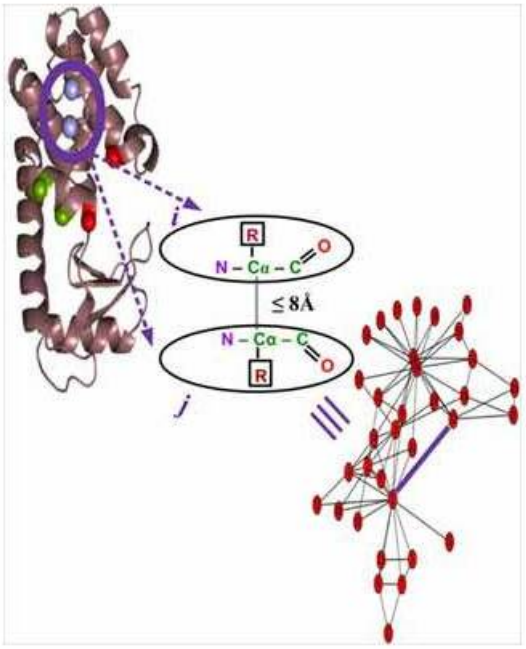

Figure 2.4. PPI network where the interaction between proteins i and j is shown as a thick edge in the graph (source- http://benthamscience.com/journal /abstracts.php, accesesed on 03/02/2016)

Analyzing this network in terms of groups of similar proteins narrows down the problem to identifying protein complexes. These protein complexes can be further explored by biologists to derive meaningful conclusions such as predicting the functions of unknown proteins and tracing the pathways involved during various activities. It can also be used in drug design area by analyzing the effects of drug administration on certain proteins and their role in causing disease.

2.1.3 Gene Expression Data

DNA in the living cell consists of segments which carry information for the synthesis of certain functional products. These segments may be coding or non-coding in nature. The coding segments, called the genes are responsible for the production of proteins while the non-coding ones like transfer RNA and small nuclear RNA are responsible for the production of functional RNA. Gene expression is therefore, the process by which encrypted information from genes is decrypted giving rise to certain proteins or other functional products. This is the most fundamental process used by all living organisms (both single and multi-cellular) to accomplish their day-to-day activities. The expression properties of these genes give rise to

specific traits in each one of us and is also responsible for catalyzing different metabolic activities. However, this is a very regulated process and is involved in monitoring the production of different substances in the body. For example, a proper expression of the insulin protein in the body monitors the blood sugar level. The expression of genes is also dependent on external factors in addition to internal ones. Cells respond to external conditions by regulating the timing and amount of functional products to be produced. This close monitoring of cells aids the existence of living beings on this earth.

2.1.3.1 Gene Expression Data Measurement Techniques

Recording the expression levels of genes can serve various functions. For example, expression of certain genes in the human body can be used to determine if a person is more susceptible to any kind of disease or how his/her body reacts to drugs used in treatment. A number of factors are responsible for regulating the process of expression of these genes. These regulatory mechanisms decide the amount of mRNA to be produced from genes at the proper place and time. There are a number of techniques to quantify the mRNA levels of genes. A few such techniques are discussed here.

- Northern Blotting : In this process, the total RNA whose expression is to be measured is first extracted from the cell. From this RNA, mRNA with a poly(A) tail is isolated using oligo cellulose chromatography. Samples of RNA are then separated using agarose gel and then transferred to a nylon membrane. A labeled probe is then hybridized with the RNA in membrane. This hybrid signals are then detected using an autoradiograph [138].

- RT-qPCR : This is a two step process involving reverse transcription followed by quantitative PCR [32]. Using the reverse transcription process, a single stranded DNA called cDNA is generated from mRNA. As the DNA amplification process progresses, the hybridization probes emit varying fluorescence. Using a standard curve, qPCR can determine the absolute number of copies of mRNA per cell.

- Hybridized Microarray : For analyzing multiple genes at the same time, a DNA microarray technology is used [107]. A DNA microarray consists of a large number of tiny DNA spots on a solid surface. Each spot contains picomoles of a specific DNA sequence. It is based on the complementary pairing of DNA. When two DNA strands are hybridized, complementary pairing takes

place. The higher the number of base pairs in a sequence, the stronger the bond between two strands. The next step involves washing off this mixture to remove the weak base pairs. A microarray chip with DNA probes is then dipped into a fluorescent dye solution. The strength of the deposition of dye at any point depends on the amount of complementary samples present on that spot. The color intensity of this dye is then converted into numeric values, which are the absolute expression values of the corresponding genes.

- RNA-Seq : It is the most recent type of technique to measure the quantity of mRNA in any sample at a given time point [16]. In addition to m-RNA quantification, it can also take care of miRNA and tRNA. In this technique, the mRNA molecules are first divided into small groups of nucleotide bases, which are then aligned with the gene's nucleotide sequence. The strength of the complementary pairing occurring between the two sequences are measured as the gene's expression level.

2.1.3.2 Types of gene expression data

Usually all gene expression technologies work on the intensity of image obtained from the mRNA level of genes. The conversion of these images into numeric values is again a difficult task and is highly platform dependent. Once these images are in numeric form, they are represented as a matrix, where rows represent the gene names and columns are the samples or the time points. Depending on the columns in this matrix, the dataset is divided into two types.

- Gene Sample Expression Dataset : A gene is said to express itself under different conditions in different subjects. This type of dataset consists of expression levels of genes for different samples. Rows in this matrix represent genes and columns are the sample names, while the value corresponding to each entry in the matrix is the expression value of the gene for that sample.

- Gene Time Dataset : This type of dataset contains the expression values of genes at different time points. Genes in a cell express themselves only at certain specific time points. If all genes were to express at the same time, monitoring their expression levels would be a very tedious process. A gene expresses itself only during certain periods. The columns in this type of dataset consist of time points and the values represent the expression levels of genes at these time points for the same sample.

These gene expression datasets are stored and maintained by institutes. They are also responsible for the timely updation of these datasets. A few well known repositories known in the literature are ArrayExpress, Gene Expression Omnibus-NCBI [9] etc.

2.1.3.3 Gene Expression Data Analysis

A gene expression dataset not only gives the amount by which genes express themselves, but also highlight other important information present in the living body. Therefore, a careful analysis of such data helps unfold the mystery behind topics such as unrevealing the regulatory links between genes, predicting the function of uncharacterized genes and association of genes with diseases [14]. Using data mining approaches of clustering and classification on these data, one can find groups of similar genes. Once a set of similar genes is found, the functionality of an uncharacterized gene can be attributed to the functionality of other members in the set. An extension of this approach would help derive regulatory relations among the genes. Gene expression data also assist in finding co-regulated and differentially expressed genes. The differentially expressed genes can be extended to identify consensus modules in a three stage disease dataset. With the help of these consensus modules, one can further extend the analysis process in different directions. One can also use the gene expression data to derive relationships between a disease and the corresponding drug administered to that patient till date. A detailed analysis of the change in expression values would reveal the patient's response to the drug, thereby assisting the drug designers to further improve the drug.

2.1.4 Data Mining in PPI and GE Analysis

The day to-day increase in the availability of biological data has given way to many challenges and opportunities. Researchers are continuously working towards understanding these data and deriving meaninful conclusions. Here comes the role of data mining, which deals with extracting relationships in these large datasets. Relationships can be in the form of correlations or patterns or any other kind of similarity among data elements. The main tasks involved in data mining and their utility in PPI and GE data analysis is discussed next.

- Regression Analysis : Regression is the science of estimating relationships among variables. It helps identify the dependency of a variable of interest with other varying parameters. This statistical technique has been ex-

plored by researchers in analyzing the expression level of genes. Li et al [82] proposed RACER (Regression Analysis of Combined Expression Regulation) which predicts the mRNA expression level as a function of copy number variation (CNV), DNA methylation (DM), transcription factor (TF) and microNA (miRNA). This has been validated using an Acute Myeloid Leukemia (AML) dataset and has shown good performance in of detecting miRNA/TF targets. Regression has been successfully used in case of PPI dataset to predict protein complexes. In [167], a regression model is trained using the properties of both weighted and unweighted PPI networks. It then finally predicts candidate complexes based on a clique filtering approach.

- Classification : Classification is the process of assigning labels to new data based on apriori classified. A model learns a classification function based on feature vectors of the apriori data (training samples) and then uses the function to assign new observations to the most similar class label. This technique can be used for PPI data as well as gene expression data. It has been widely used in analyzing gene expression data and deriving conclusions on its benign or malicious nature. Various studies [106] have analyzed the performance of existing classification techniques such as SVM, RBF, Bayesian and Decision Trees on microarray gene expression datasets.

- Clustering : This is the most commonly used approach for large biological datasets. It is the grouping of data in such a way that intra-group similarity is high and inter-group similarity is low. Similarity is computed differently by different researchers giving rise to a number of clustering algorithms. For biological data, the major ideas in computing similarity are distance between two elements in the network or biological similarity between the two elements. A few established clustering algorithms on gene expression data are Hierarchical Clustering (HC), Self-Organizing Map and Self Organizing Tree Agorithm [165]. MCODE, DECAFF, RNSC, ClusterONE and GMFTP [130] are a few well-known clustering algorithms available for PPI network.

- Association rule mining : It is the process of estimating relations among elements that take numeric values. It is based on the concept of an antecedent, which is identified in the dataset and a consequent, which is found to co-occur with the antecedent. This technique has been used to widely handle the missing nature of microarray data. It has also been used to predict links between proteins in a PPI network [54].

2.1.5 PPI and Gene Expression Data Analysis Tools

The work reported here uses a number of platforms and tools. I have used these platforms to code different techniques. The use of tools also has been an integral part of my research. These tools have allowed me to do away with some existing implementations. I have also used these tools to derive certain results using their interface. I now discuss them in detail.

- MATLAB : MATLAB (Matrix Laboratory) [91] is a part of the fourth generation programming language and environment. It is proprietary software developed by the MathWorks, which is an American company specializing in making mathematical tools. It provides the user with many built-in functions for common use. It is based on matrix operations and provides certain other functions such as graph plotting, user interface creation and facility for interfacing with other programming languages such as C, C++, Java and Python. Another added advantage provided by Matlab is that it supports collaboration with third parties. For example, if one wishes to use an algorithm for signals processing, one can avail oneself of the benefits of the *Signal Processing Toolbox* in Matlab. Similarly, if a biologist wants to use an alignment algorithm, he can avoid coding the whole algorithm by using the *Bioinformatics Toolbox*. The newer versions of Matlab can make use of multicore processors in a system by using the *Parallel Computing Toolbox* [128]. This is one of the most attractive features that Matlab has incorporated and is being used on a large scale. I used the Parallel Computing Toolbox to get faster execution of my algorithms. In one case, I had to process around 50,000 entries which was taking more than 10 days for sequential execution. I used the parallel concept and divided my processing into worker nodes and it was solved in an hour. This is the power of the Parallel Computing Toolbox.

- R package : R [143] is an open source programming language and is widely used in the research community because of its easy availability. Its superiority is based on the availability of a set of extensive packages available under its umbrella. The source code of R is written in C, Fortran or R itself, and hence the coding is like C programming. Another feature of R is that it is much faster than MATLAB and C. I have used the *GOSemSim* [168] package available under R to find the semantic similarity between a set of gene products. This package provides easy implementation of many types of semantic similarity measures such as Resnik, Lin and Jiang. However,

despite of its ability to compute results at a faster rate, it is not very popular. This is because of its package dependencies. In order to make one package workable, other packages need to be installed. These installed packages have to be compatible with the R version in use, otherwise the package installation would fail.

- Cytoscape : Cytoscape [126] is an open source Java based software available for visualizing large biological networks. Although it can be used for other types of networks too, it is most widely used in the biological domain for PPI networks, gene expression networks, disease networks, integrating biomolecular entities to certain networks etc. It also has an extensive library of plugins available for various purposes. I have used *ClusterViz, CytoCluster, CommFinder, ClusterONE, ClustnSee* plugins of cytoscape for protein complex finding in PPI networks. I have also used *BinGO* [89] plugin to find overrepresented GO terms among a set of genes/proteins. This is very useful stand-alone tool to find p-value of a set of genes.

- ProCope: ProCope [71] is a Java based tool which has certain complex finding techniques such as MCL and HAC implemented in it. It also provides a GUI to evaluate the effectiveness of complexes against a set of bonafide complexes. I have used this tool to calculate the performance (in terms of Positive Predictive Value, Sensitivity and Accuracy) of my complex detection techniques just by using the requisite benchmark set.

- DAGO-Fun : DAGO-Fun [93] is an online tool which is used to find the semantic similarity between GO terms associated with proteins. Semantic similarity determines the relationship among the GO terms based on its position in the Directed Acyclic Graph. This tool implements both annotation and topology based semantic similarity measures. I have used it to find Wang's semantic similarity between a set of proteins. However, it has a few limitations- - it requires good internet connectivity and the maximum number of proteins in each turn cannot exceed 2000.

- GeneAnnot: GeneAnnot [30] is a web based tool maintained at the Weizmann Institute of Science. It has a GUI which provides ID conversion from HG-U95, HG-U133 and HG-U133 Plus 2.0 to gene symbols or gene IDs. I have used this tool to get the gene names corresponding to Affymetrix IDs, HG-U133.

- DAVID : The Database for Annotation, Visualization and Integrated Discovery (DAVID) [25] is a multi-purpose tool, which provides many services such

as ID conversions, functional enrichment, gene-disease associations, pathways information etc. I have used this tool to identify various pathways to which specific genes are linked.

- KEGG Pathway database: The KEGG database [61] provides a GUI for specifying the user's query . It gives a visual layout of the different terms, genes and pathways associated with the queried structure. I have used it to find the biochemical pathway associated with certain diseases.

2.1.6 Datasets Used

Various datasets are available to validate the performance of existing data mining techniques in the field of PPI data analysis and gene expression data analysis. In this work, I have focussed mainly on PPI analysis and disease analysis from the view point of gene expression. Therefore, I have used only these two types of datasets. Details of these datasets are discussed next.

2.1.6.1 PPI dataset

The PPI datasets represent interactions among proteins occurring in an organism. These interactions are recorded by various experiments such as Mass Spectrometry, Tandem Affinity Purification etc. These datasets are represented as a set of pairs of proteins delimited by the tab character. Certain PPI datasets highlighting the weights of these pair of interactions are also available, but I have restricted my thesis to unweighted datasets only. The PPI datasets used in my work are listed here. The first five datasets are of yeast and the last one is of human.

- Gavin_2002 : Researchers started analyzing multi-protein complexes in Saccharomyces cerevisiae using large scale experimental techniques of mass spectrometry and tandem affinity purification [35]. During this process, they started recording the interaction status and interacting partners of each protein. This recorded version is what is called Gavin_2002 dataset today. It comprises of 1352 proteins and 3210 interactions.

- Gavin_2006 : A genome-wide screening through mass spectrometry of complexes in budding yeast gave rise to this dataset [36]. It consists of 1430 proteins and 6531 interactions.

- Krogan_2006: This dataset is prepared by rigorous analysis of tagged proteins in yeast. Analysis of these proteins is performed by both ionization in mass

spectrometry and liquid chromatography in tandem mass spectrometry [70]. It consists of 2675 proteins and 7088 interactions.

- Tong_2004: This dataset is created by mapping cross mutations in genes to a viable set of gene yeast deletion mutants [144]. It consists of 2262 proteins and 7430 interactions.

- DIP: The Database of Interacting Proteins (DIP) is actually a database which contains information about proteins found in different organisms. They store the protein name, its sources of experimental information, information on individual experiments, interaction partners of each protein, etc [122]. The interaction information of each protein is also available in a pair-wise format, which is the DIP dataset. It consists of 4930 proteins and 17201 interactions.

- HPRD : The Human Protein Reference Database is a repository which contains information about interactions among proteins, changes occurring in them after translation, enzyme-substrate relation and additional other such informations [63]. Information about interactions among proteins in humans obtained via yeast-two-hybrid, in-vivo and in-vitro methods. This dataset consists of 39240 interactions among 10080 proteins.

A tabular detail of these datasets is given in Table 2.2.

Table 2.2 Datasets Used

Organism	**Name**	**Number of proteins**	**Number of interactions**	**Means of preparing the dataset**	**Availability**
Yeast	Gavin_2002	1352	3210	Tandem affinity purification and mass spectrometry.	*http://www.bioacademy.gr/bioinformatics/projects/GIBA/*
	Gavin_2006	1430	6531	Genome wide screening test.	
	Krogan_2006	2675	7088	Tandem affinity purification, mass spectrometry and machine learning methods.	
	Tong_2004	2262	7430	Cross mutation and tandem affinity purification.	
	DIP	4930	17201	Uses a number of resources to derive interaction pairs	*http://dip.mbi.ucla.edu /dip/page?id=download*
Human	HPRD	10080	39240	Curated database	*http://www.hprd.org/*

2.1.6.2 Gene Expression Disease datasets

GSE or the Genomic Spatial Event is repository that stores all types of microarray data. It handles expression data, functional annotation data, genomic annotations data, etc. [23]. This repository is maintained by NCBI. For my research work, I have used disease gene expression data, which are available in this repository. These dataset comprises of expression values of 22,283 genes for different subjects. These subjects are nothing but individuals for which the expression level of these genes are recorded. I now discuss the two GSE datasets used in my work.

- GSE 2034 : This dataset is called the *Breast Cancer Relapse Free Survival Dataset*. It consists of expression levels of genes in non-metastasis and metastasis stages of breast cancer. The dataset comprises of 286 samples out of which 180 samples are of non-metastasis stage while the rest 106 show expression levels of genes during metastasis stage. The platform used for measuring the gene expression is GPL96 [HG-U133A] Affymetrix Human Genome U133A Array. This dataset was created at Veridex in San Diego, USA. I used the most updated version of the dataset, which was last updated in July, 2016. The link to the dataset is *https://www.ncbi.nlm.nih.gov/geo/query /acc.cgi?acc=GSE2034.*

- GSE 8397 : This dataset is called the as *Expression Profiling of the Parkinsonian Brain*. It consists of expression level of 22283 genes over 47 samples. Out of these 47 samples, 23 are from control state and the other 24 are diseased samples. The platform used is GPL96 [HG-U133A] Affymetrix Human genome U133A and GPL97 [HG-U133B] Affymetrix Human genome U133B. The dataset was created at Imperial College, London, United Kingdom. It has been last updated in July 2016 and I have used this version of the dataset in my work. This dataset is available at *https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi.*

2.1.7 Biological Knowledge bases

A biological knowledge base is a store-house that contains biological expertise information based on findings of human experts. These knowledge bases can be used to validate different computational outputs. I have made use of some available information sources to validate my computational results.

2.1.7.1 PPI Gold Standards

There are certain repositories which maintain quality information about the interactions among proteins. These interactions are derived from high-throughput experiments and are nearly free of false positive data. Such datasets are often used as benchmarks to evaluate the quality of the computed results. I now discuss the three gold standards (benchmark) datasets used in my research work.

- MIPS : The Munich Information Center for Protein Sequences (MIPS) [95] maintains a database of high quality genome information derived from several rigorous experiments. It maintains genome information for a number of organisms such as yeast, Arabidopsis thaliana and Neurospora crassa along with protein sequence information. I used the yeast PPI information available in this repository. This benchmark set consist of 203 complexes and was prepared in May, 2006. It can be downloaded from *http://www.paccanarolab.org/cluster-one/*.

- CYC2008 : The CYC2008 benchmark [112] set comprises of 408 complexes in yeast, which are derived from small-scale experiments and have direct links to published literature. It is available at *http://wodaklab.org/cyc2008/*.

- PCDq :PCDq [64] is a repository which includes human protein complexes along with quality indices derived from both known and predicted complexes. This complex dataset was built using a combination of six PPI repositories-DIP, MINT, BIND, HPRD, GNP_Y2H and IntAct and information from predicted complexes from PPI networks. This comprises of 1264 protein complexes and can be downloaded from *http://h-invitational.jp/hinv/pcdq/*.

2.1.7.2 Co-localization Datasets

Colocalization information predicts the location of each element in the cell. Usually proteins in a complex belong to the same cellular location. One can use this information to evaluate the localization score of the predicted complexes. I use the *Kumar [53]* and *Huh [72]* colocalization datasets provided by ProCope to measure the effectiveness of my method. However, I restricted this index only to yeast dataset as there was no ready to use colocalization dataset available for humans.

2.1.7.3 Gene Ontology

Gene Ontology [3] is an initiative by the Bioinformatics department to unambiguously represent genes as well as gene products across all species. This project annotate genes and its products with various terms technically referred to as GO terms. It also makes note of the relationships that exists between these terms in the GO hierarchy. It stores the biological, molecular and cellular role of genes and their products in its structure. I used the Gene Ontology concept at the back end to find p-value and semantic similarity between protein pairs using the BinGO plugin in Cytoscape and DAGO-Fun tool respectively.

2.1.7.4 GeneCard

GeneCard [120] is a repository of human genes maintained at the Weizmann Institute of Science, Israel. It comprises of the genomic, proteomic and functional information of all known genes. It provides information about more than 7000 human genes from more than 90 different sources. I have used GeneCard for finding the list of genes associated with Alzheimer's Disease, Breast cancer and Parkinson's Disease.

2.1.8 Data cleaning

The quality of data used in any analysis plays a very significant role in deciding the results. Biological data are prone to be noisy and redundant in nature. Such kind of data leads to poorer analysis. Therefore, one has to get rid of such data before jumping into any outputs. In this work, we have not encountered much noisy and missing data as we are using data from curated databases. However, there are situations where we find redundant data and we got rid of such anomalies before using them for our analysis. In some of the works proposed here, we have used certain other steps like normalization and discretization of data which have been discussed as and when used.

2.1.9 Performance Indices

Protein complexes are groups of proteins which are found using clustering techniques on the PPI data. In order to evaluate the performance of our complex finding methods, these complexes have to be matched against a set of benchmark complexes. Details of the benchmark complexes are given in Subsection 2.1.7.1. It

is obvious that complexes detected using any technique would never exactly match the gold standard set. Therefore, there is a need for a metric to measure matching overlap, accepted by the research community, to consider a predicted cluster as complex. This is known as the overlapping threshold. Two overlapping schemes have been proposed in the literature [131]. These are known as Bader's scheme and Wang's scheme.

Suppose a cluster obtained computationally consists of $n_{cluster}$ proteins and a benchmark or true set consists of $n_{complex}$ proteins. Let us assume that n_{common} is the number of common proteins between the sets. Then the overlapping scores are defined by Equation 2.1 and 2.2 respectively.

$$Ov_{Bader} = \frac{n_{common}^2}{n_{cluster} \times n_{complex}} \tag{2.1}$$

$$Ov_{Wang} = 2 \times \frac{n_{common}}{n_{cluster} + n_{complex}} \tag{2.2}$$

For analysis, researchers have fixed standard values for these overlapping scores, $Ov_{Bader} = 0.2$ and $Ov_{Wang} = 0.6$. A cluster is said to match a benchmark complex only if its overlapping score is greater than or equal to the standard value. This matching of predicted cluster with the complex set is then used to calculate the Precision, Recall and F-measure of our methods.

2.1.9.1 Precision, Recall and F-measure

Precision, Recall and F-measure are used to evaluate the quality of complexes obtained by a complex finding process. *Precision* is the percentage of match between the predicted clusters and the complexes [131]. *Recall* on the other hand obtains the fraction of known complexes that has been detected within the predicted clusters using the complex finding method.

Suppose a complex finding method predicts P_C clusters. The performance of this method is validated against a gold standard consisting of B_C complexes. Using the overlapping score, the match between predicted clusters and complexes is found. Let N_{pc} be the number of predicted clusters that match at least one benchmark complex and N_{bc} is the number of benchmark complexes that match atleast one predicted cluster. Then the Precision and Recall of this method are

given by Equations 2.3 and 2.4, respectively.

$$Precision = \frac{N_{pc}}{P_C} \tag{2.3}$$

$$Recall = \frac{N_{bc}}{B_C} \tag{2.4}$$

Another index, called F-measure, which is the harmonic mean of the Precision and Recall can be used to evaluate the effectiveness of the method. F-measure is given by Equations 2.5.

$$F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{2.5}$$

A good complex finding method should have high precision and recall values so as to ensure a high F-measure. A high precision value indicates better coverage of the complex finding method considering the set of benchmark complexes.

Due to the increasing availability of knowledge of genomes, some researchers are of the opinion that overlapping score threshold should be increased to 0.75 for the yeast dataset. However, existence of false positives in the human PPI dataset makes it impossible to use this threshold for this dataset. This inconsistency observed during evaluation of the effectiveness of the method can be handled using other means, which are free from such thresholding. This leads to the use of Positive Predictive Value, Sensitivity and Accuracy to determine the performance of these methods.

2.1.9.2 Positive Predictive Value, Sensitivity and Accuracy

I tried to validate the performance of my complex finding methods over yeast as well as human PPI datasets. Thus, to avoid bias while using the overlapping threshold, I use Positive Predictive Value, Sensitivity and Accuracy measures. These indices can be understood with the help of a cross-tabulation matrix, X, which is of the order $P_C \times B_C$, where, P_C represents the number of predicted complexes and B_C represents the number of benchmark complexes. Each entry of the matrix, $n_{common_{ij}}$ represents the common proteins between the benchmark, i and predicted cluster j. Using this matrix, we now define these indices.

Positive Predictive Value : It is a measure of how much the predicted cluster set matches that of the benchmark set. It is calculated for every cluster set as $PPV_{cluster_j} = max_{i=1}^{P_C} PPV_{ij}$, X_j being the marginal sum of the j^{th} cluster. For a

cluster, PPV represents how closely a predicted cluster resembles its best matching complex. To compute the PPV over all cluster sets w.r.t. all annotated complex, it is desirable to get an overall PPV. The overall PPV of a method is the average PPV's of all clusters, i.e.,

$$PPV = \frac{\sum_{j=1}^{P_C} X_j.PPV_{cluster_j}}{\sum_{j=1}^{P_C} X_j} \tag{2.6}$$

A high PPV value indicates higher fraction of correspondence between the predicted cluster and complex, which indirectly implies better quality results.

Sensitivity : It computes the fraction giving how much of the benchmark set is contained in the predicted cluster set. Sensitivity of a complex, Sn_{cmplx_j} is defined as the maximum value of sensitivity obtained for complex j over all B_C real complexes. Mathematically, it is given as $Sn_{cmplx_j} = max_{j=1}^{B_C} Sn_{i,j}$ where $Sn_{i,j} = \frac{X_{i,j}}{N_i}$, N_i represents the cardinality of complex i. The overall sensitivity is the weighted average of the individual ones and is defined as.

$$Sn = \frac{\sum_{i=1}^{B_C} N_i Sn_{cmplx_i}}{\sum_{i=1}^{B_C} N_i} \tag{2.7}$$

Higher sensitivity of a method indicates larger coverage of the predicted clusters by the benchmark complexes. A good complex finding method requires a trade-off between PPV and Sn. Accuracy is defined to establish this trade-off.

Accuracy: To get a compromised yet effective value, considering both PPV and Sn, the geometric mean of the two defines accuracy of the method. Mathematically,

$$Acc = \sqrt{PPV \times Sn} \tag{2.8}$$

Accuracy can be used as an unbiased measure to evaluate the statistical effectiveness of any complex finding method with the help of gold standards. To analyze the predicted clusters from a biological point of view, the standard Gene Ontology repository is used. Details of the Gene Ontology are discussed in Subsection 2.1.7.3.

2.1.9.3 Co-localization score

The Co-localization score is effectively used to evaluate the predicted complexes w.r.t. a standard localization dataset. If V_p is the total number of proteins in cluster C_i which reside in location p and V is the total number of proteins in C_i,

then co-localization score of C_i is computed as follows.

$$Colz(C_i) = max(\frac{V_p}{V}) \tag{2.9}$$

2.1.9.4 p-value

In simple terms, $p-value$ indicates how strongly results are supported by a knowledge base. It gives the probability by which a given set of proteins is enriched by a functional group, G, by chance. Mathematically,

$$p - value = 1 - \sum_{i=0}^{N_i - 1} \frac{\begin{pmatrix} |G| \\ i \end{pmatrix} \begin{pmatrix} |V| - |G| \\ |C_i| - i \end{pmatrix}}{\begin{pmatrix} |V| \\ |C_i| \end{pmatrix}} \tag{2.10}$$

where C_i is the predicted cluster containing N_i proteins in G, and the entire PPI network contains $|V|$ proteins.

2.1.10 Discussion

Protein protein interaction data and gene expression data have been widely studied to comprehend the mystery behind the existence of living beings. To handle this large and diverse data, various computational techniques have been devised. The output from these techniques has been effectively used in inferring biological information. I have explored certain techniques in data mining to handle a few issues associated with such data. PPI data analysis, from clustering perspective is the focus of this work. An application to this analysis has also been discussed in the form of ranking the groups obtained from PPI network w.r.t. diseases. I have also explored the possibilities of finding modules from gene expression data. This part of the work has also been extended towards the study of progression of diseases.

In the next chapter, I have handled the PPI data analysis part in terms of protein complex finding and proposed two methods -CNCM and DCRS. The methods have been validated on real datasets- yeast and human PPI dataset.