DATA MINING AND NETWORK MINING APPROACH FOR SIGNIFICANT PROTEIN IDENTIFICATION IN GASTRIC BYPASS SURGERY

A Thesis in
Industrial Engineering
by
Manini Madireddy

© 2009 Manini Madireddy

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

May 2009
The thesis of Manini Madireddy was reviewed and approved* by the following:

Soundar R.T. Kumara
Allen E. Pearce/Allen M. Pearce Professor of Industrial Engineering
Thesis Advisor

Tao Yao
Associate Professor of Industrial Engineering

Richard J. Koubek
Peter & Angela Dal Pezzo Professor of Industrial Engineering
Head of the Department of Industrial and Manufacturing Engineering

*Signatures are on file in the Graduate School.
Abstract

In this research project our focus was to understand the underlying biological causes, identify the significant proteins and the most probable drug targets of diabetes. In the process of doing this we will be developing a generic methodology of significant protein identification for any disease. There are a number of oral anti diabetic drugs available in the market but none of them have worked effectively towards prevention and cure of the epidemic diabetes; none the less have adverse side effects. The reason for this is the lack of complete understanding of diabetes due to its complexity and the multifactorial components involved in disease onset and progression.

A disease or a biological function can only rarely be attributed to a molecule - it is usually an interaction of numerous molecules that perform a function or bring about any change. So one just cannot look at the genes or proteins or metabolites; but needs to look at all of them and their interactions to come up with a good drug target. This is a more holistic approach of understanding disease and is called the systems biology based approach. A natural way to represent such
discrete interacting units is a graph or network.

Differentially expressed protein ratios of morbidly obese patients before and after bariatrics surgery is obtained by researchers at the Hershey Medical Center which was used in this work. Surprisingly, some of the patients, who were diabetic before the surgery, were freed from diabetes. This observation led us to the hypothesis that the proteins expressed in the patients may be related to the potential drug targets for diabetes. In order to identify drug targets, we adopt a systems biology approach i.e. analyze the proteins observed from patients in the context of the entire human protein interactions. We believe that this approach will help us identify key proteins that play a major role in diabetes.
Table of Contents

List of Figures viii
List of Tables ix
List of Symbols x
Acknowledgments xi

Chapter 1
Introduction 1
1.1 Problem Definition ...................................... 1
1.2 Background .............................................. 2
  1.2.1 Diabetes ........................................ 4
  1.2.2 Bariatric Surgery .................................. 6

Chapter 2
Healthcare and Systems Biology 8
  2.1 Introduction ........................................... 8
  2.2 Healthcare - Phases of Development ...................... 9
    2.2.1 Description of the phases .......................... 10
    2.2.2 Systems Biology in Healthcare ...................... 11
  2.3 Drug Discovery ........................................ 13
    2.3.1 Impact of Systems Biology in Drug Discovery ....... 14
  2.4 Digital Biology ........................................ 15
    2.4.1 Scientific data integration ........................ 16
      2.4.1.1 Modeling ...................................... 18
      2.4.1.2 Simulation and Analysis ....................... 18
    2.4.2 Multi-scale Biological modeling .................... 19
List of Figures

3.1 Mass Spectrometer and its components. ....................... 25

4.1 Plot of Class 1 Proteins - Average Ratio and Frequency. ........ 40
4.2 Plot of Class 2 Proteins - Average Ratio and Frequency. ........ 41
4.3 Plot of Class 3 Proteins - Average Ratio and Frequency. ........ 41
4.4 Proteins - Resistant Pareto Front. ................................ 42

5.1 Sample Transcription Regulatory Network ..................... 64
5.2 Sample Protein Interaction Network ............................... 65
5.3 Sample Metabolic Network ........................................ 66
5.4 Protein Interaction Network ...................................... 68
5.5 Degree Distribution. ........................................... 69
5.6 Network Diagram-Betweenness Centrality ....................... 70
5.7 Betweenness Distribution. ........................................ 71
5.8 Clustering Distribution. .......................................... 72
5.9 Protein Interaction Network showing the different modules. ....... 73
List of Tables

3.1 Comparison table for Mass-Analyzers ................. 25
4.1 Proteins of interest ................................... 43
5.1 Summarized Network Parameters .......................... 75
5.2 Top 15 Proteins of interest ............................... 76
6.1 Top 15 Proteins of interest ............................... 79
List of Symbols

\( \bar{r}_i \)  Mean of a protein ratio across all the patients, p. 25
\( r_{ij} \)  The ratio of protein j in the ith patient, p.25
\( \eta_i \)  Number of patients in which protein i is present, p.25
\( \sigma_i \)  Standard Deviation of Protein i across all the patient, p.25
Acknowledgments

I thank my advisor Dr. Soundar Kumara for his continuous support during the Masters' program. He has been a mentor, grooming me to scientific writing and presentations, and encouraging a quantitative approach to problem solving. I express my gratitude to my thesis reader Dr. Tao Yao for helping me complete my thesis in time. I am deeply indebted to Dr. Bruce Stanley and his research group at Hershey Medical Center for providing us with invaluable data. I appreciate Dr. Debashis Ghosh of the department of Bio-Statistics for spending time in understanding the problem and providing me with an initial direction.

I am grateful to my peers in the Laboratory of intelligent System and Quality (LISQ) research group who patiently helped me formalize my thesis problem over repeated meetings.

I am deeply indebted to my parents who were a source of motivation and strength thought my career. I am grateful to my friends for standing by me through the best and worst of times, and to the Almighty for aligning the turn of events in my good fortune. I wish to acknowledge the National Institute of Standards and Technology for their generous financial support for making this work possible under the Grant number 60NANB6D6205.
Dedication

To my parents.
Chapter 1

Introduction

1.1 Problem Definition

The mass spectroscopy group headed by Dr. Bruce Stanley at the Hershey Medical Center discovered that few of the diabetic morbidly obese patients who underwent bariatric surgery (more popularly known as weight loss surgery) found themselves freed from diabetes after the surgery. Technological advancements have made it possible to capture the "snapshot" of the biological state of the patient before and after the surgery in terms of the relative protein levels - Differential Protein Expression Profiling.

The objective of this thesis is to understand and identify the biological elements, in this case proteins; that brought about the transition of the patients state from being diabetic to a state of being non-diabetic with respect to the bariatric surgery. This is a challenging problem because of the following issues:
1. There are six patients, and each individual’s genetic makeup is different. Implying different levels of proteins get expressed. This will lead to some variation in the way his or her body responds to the surgery.

2. The human body is a complex adaptive system and there are a large number of interactions between its components. Hence, it is not enough to identify proteins involved in the transition; we also need to identify interactions of proteins.

1.2 Background

Dr. Bruce Stanley’s group at Hershey Medical Center collected biological samples (containing proteomic information) from the diabetic morbidly obese patients who underwent gastric bypass (a type of bariatrics surgery). Some of these patients, who were diabetic before the surgery, were freed from diabetes after the surgery. This observation led us to the hypothesis that these proteins expressed in the patients may be related to the potential drug targets for diabetes. There are a number of oral anti diabetic drugs available in the market but none of them have worked effectively towards prevention and cure of the epidemic diabetes. The reason for this is the lack of complete understanding of diabetes due to its complexity and the multifactorial components involved in disease onset and progression [1]. Most of the current therapies for diabetes were developed in the absence of drug
targets [2]. The first few diabetic therapies were focused towards the reduction of hypoglycemia and the current therapies aim to sensitize insulin secretion and to inhibit hepatic glucose output [3]. There are currently six different classes of hypoglycemic agent being used- insulin, sulfonylureas, meglitinides, biguanides, alpha-glucosidase inhibitors and thiazolidinediones [4]. All these drugs were developed without taking the holistic approach into account and i.e. why most of these drugs are not a cure and have adverse side effects.

One just cannot look at the genes or proteins or metabolites and identify the drug targets. One needs to generate a database of all the factors that affect diabetes and then look for drug targets i.e. one needs to look at the genes, proteins and the metabolites as well as their interactions to identify the drug targets [1]. This is a more holistic approach of understanding disease and is called the systems biology based approach. The systems biology approach involves computation, experimentation and enquiry; this is an ideal approach for drug discovery and personalized medicine.

Diabetes is the major cause of morbidity and mortality in the U.S. [5]. There are 230 million diabetics around the world according to 2006 statistics [6] and expected to reach 380 million by 2025 [7]. Thus, disease pathway and drug targets identification are the inventions of high priority as diabetes is a fast growing epidemic. The most important epidemic of the world is overweight. There are 1.7 billion over weight individuals in the world of which more than half are obese. The
four most important comorbid conditions are - diabetes, hyperlipidemia, hypertension and obstructive sleep apnea. Diabetes is one of the most common comorbid conditions of obesity.

Diabetes is a metabolic disorder characterized by hyperglycemia (high blood sugar level). The World Health Organization (WHO) recognizes three main types of diabetes; they are Type 1, Type 2 and Gestational diabetes (this is a temporal condition when one is diabetic only during pregnancy). All these types of diabetes have similar signs, symptoms and consequences but they are different in their pathophysiology (the study of biological function during states of disease or dysfunction).

1.2.1 Diabetes

Diabetes is a metabolic disorder where the pancreatic $\beta$ cells either produce little insulin or completely cease to produce insulin; or the receptors on the cells of the human body progressively grow resistant to insulin. It is also defined as a state in which lipids and carbohydrates are improperly regulated by insulin [8]. As already mentioned the World Health Organization recognizes three main forms of diabetes: type 1, type 2, and gestational diabetes (occurring only during pregnancy), which have similar signs, symptoms, and consequences, but different causes and population distributions. Diabetes is a syndrome of a group of diseases (rather than one disease), leading to the prolonged hyperglycemic state. Ultimately, all forms
are due to the disturbed balance between the two opposing hormones insulin and glucagon.

Type 1 is a *autoimmune disease* that results in the permanent destruction of the pancreatic beta cells which produce insulin. This form of diabetes accounts for 10 percent of all the cases. Type 2 is characterized by tissue-wide insulin resistance (the normal amount of insulin produced in the body will not be sufficient) and varies widely; it sometimes progresses to loss of beta cell function [9]. Type 2 diabetes is also referred to as *insulin resistance* or *metabolic disorder*. The receptors on cells in the body that normally respond to the action of insulin fail to be stimulated by it - this is known as insulin resistance. In response to this more insulin may be produced, and this over-production exhausts the insulin-manufacturing cells in the pancreas. This type of diabetes accounts for 90 percent of the cases and this number is raising. Gestational diabetes is similar to type 2 diabetes, it involves insulin resistance; the hormones of pregnancy cause insulin resistance in those women genetically predisposed to developing this condition.

Types 1 and 2 are incurable chronic conditions, but have been treatable since insulin became medically available in 1921, and today they are usually managed with a combination of dietary treatment, drugs (in type 2), and insulin supplementation. Gestational diabetes typically resolves with delivery. Diabetes can cause many complications [10]. Acute complications like hypoglycemia (is a medical term referring to a pathological state produced by a lower than normal level of sugar in
the blood), ketoacidosis (which is caused by high concentrations of ketone bodies, formed by the deamination of amino acids, and the breakdown of fatty acids) or nonketotic (is a type of diabetic coma associated with a high mortality seen in diabetes mellitus type 2) may occur if the disease is not adequately controlled. Serious long-term complications include cardiovascular disease (doubled risk), chronic renal failure (diabetic nephropathy is the main cause of dialysis in developed world adults), retinal damage (which can lead to blindness and is the most significant cause of adult blindness in the non-elderly in the developed world), nerve damage (of several kinds), and microvascular damage, and poor healing.

### 1.2.2 Bariatric Surgery

Overweight is the most important epidemic of the world. Most of the morbidly obese people undergo gastric bypass (a type of bariatric surgery)[11][12]. Bariatric surgery is often referred to as a weightloss surgery. Bariatric surgery is a treatment option for the people who have not experienced a long term weight loss through other means like diet, exercise and oral therapy [12]. There are currently no truly effective pharmaceutical agents that can treat morbid obesity.

Bariatric surgery is a clinical term for many different procedures which help the patient lose weight. The common bariatics surgery procedures are - Restrictive procedure, Malabsorption procedure and a combinatory procedure [13].

Restrictive procedure is one in which the surgeon creates a small stomach pouch
which limits the amount of food patient can eat. In the malabsorption procedure the surgeon re-routes the small intestine so that food skips a portion of it. The small intestines absorb calories and nutrients from the food. Avoiding a part of it means that some calories and nutrients are not absorbed. Malabsorption procedure is usually carried out with a restrictive procedure.

The data collected by Dr. Stanley’s group at Hershey relates the data to patients who underwent gastric bypass. This is a combination procedure. In this procedure surgeon first creates a smaller stomach pouch, this helps in feeling fuller sooner [14]. Then the small intestines are re-oriented so that the absorption is reduced. There are few benefits to this, they are- resolves type 2 diabetes, high blood pressure and high cholesterol. The concerns for patients - usually suffer from dumping syndrome and need to take multi-vitamin supplements.
Healthcare and Systems Biology

2.1 Introduction

The present US healthcare system is ranked at the 37th position in the world according to the World Health Organization (WHO). The reason being, the present healthcare system focuses on treating late stage diseases or patients in a critical state, which are very difficult and expensive to treat. Prevention or an early treatment would have been a better option. The present aim is to transform the existing Healthcare system into one that is more effective in terms of treatment and finance. Many thoughtful and informed leaders of American medicine feel that the current practice of medicine, which is mainly reactive, reductionist, myopic and evidence based scientific medicine should be transformed into one i.e. Predictive, Preventive, Personalized and in the future also Participative (more commonly known as P4 medicine). The healthcare system a century ago was fo-
cused on individualized care. The main reason for this was the lack of standards. Claude Bernard at the end of the 19th century introduced scientific approach into medicine [15]. This approach was based on observations and experimental results. They called it evidence based medicine, and was focused on standardizing medical practices and had no room for idiosyncrasy [16]. The researchers of the individualized evidence based medicine are working towards identifying the biomarkers that can be used to classify the patients into subgroups. Then at a later stage, drugs can be designed for each sub-group.

2.2 Healthcare - Phases of Development

Healthcare to a large extent is an information management endeavor. There are two types of medical information, one is the patient specific data called, the *patient centric data* and the other is the general knowledge base called the *knowledge centric data*. Fragmented parts of the patient specific information is already available in the forms of personal emergency cards, mother-child record and vaccination certification. The most important aspect of personalized medicine is to gather, accumulate and manage all the patient specific data. This data can be used in future to make a knowledge base. The advantages of having such a database is discussed in the digital biology section 2.4. Consider, an example where a patient goes to see doctor with a complaint. The doctor first analyzes the patients state i.e. gets an understanding of all the symptoms. Then, recommends the patient
to undergo some tests. The doctor relates the symptoms to the test results and then prescribes a therapy i.e. medication to the patient. We call the current medical practice reactive in the sense that all the treatment is based on the reaction i.e. patient’s complaint or symptoms. The patient then undergoes some tests this will provide them with evidence. Then, the treatment prescribed by the doctor is based on the reaction and evidence provided by the patient. The term reactive in the medical context means falling sick or being sick. The current medical practice waits for the patient to fall sick and then administer the required or necessary treatment. The doctor recommends certain tests which will help him in diagnosis and suggesting therapy. The present medical system has only a few parameters that are used as indexes/metrics to measure or can be measured. Myopic means use only a few measurements to diagnose the disease and are generally unable to make finite distinctions between individuals or between subtle variations of the same disease.

2.2.1 Description of the phases

The present aim is to improve the existing Health care system into a one that is more effective in terms of treatment and finance. The researchers of today are working towards the personalized medicine. The key word for the personalized medicine phase is the patient centeredness. This phase has brought about alot of changes in the healthcare system and it seems to be effective [16]. The main aim of
this phase is to integrate the personalized medicine with the knowledge base [17].

As mentioned in the section 2.1 the medical practices have evolved over the years. The main reason for the evolution was to improve over the current practices or to add a new dimension to the existing methods of practice. The availability of large amount of data because of the advent of the new technologies was one of the main causes of the evolution of medical practices. The phases of development of medical practices is as follows [18]:

- **Reactive, myopic, evidence based, reductionist and scientific approach**: It is still the most widely used approach even today [16].

- **Predictive and Preventative approach**: In this approach the focus is on early detection of the disease and prevention.

- **P³, Predictive, Preventative and Personalized approach**: In this approach focus is on individualized medicine or patient centeredness.

- **P⁴, Predictive, Preventative, Personalized and Participative approach**: In this approach the focus is on individual’s participation in the medical practice

### 2.2.2 Systems Biology in Healthcare

Healthcare has many aspects, the most important of which are- detection and monitoring of disease, drug discovery, treatment evaluation and ultimately leading to
predictive and preventive medicine. Healthcare to a large extent is an information management endeavor. As technologies mature, they will accommodate smaller sample volumes and will be more economical, this in turn lead to personalized medicine. Systems Biology has two important roles to play in medicine. They are as follows-

1. Systems biology will continually improve the capacity to understand and model biological systems on a more global and in-depth scale than ever before. The researchers continue to gather system level insights.

2. Continual spawning of new technologies, which will enhance the efficiency, scale and precision with which cellular measurements are made. This will facilitate all aspects of healthcare.

Systems biology is used to understand the normal biological systems and the pathological states. The ability to predict and prevent diseases will always be influenced by the fundamental knowledge of the normal and diseased state of cells. Drugs and therapies can be better directed at re-engineering the behavior of malfunctioning cellular networks. The root causes of diseases can be detected by identifying the bio-markers and then by defining a treatment that influences that specific molecule. The behaviors of most biological systems, cannot be attributed to a single molecule or pathway, rather they emerge as a result of interactions at multiple levels, and among many cellular components.
2.3 Drug Discovery

In the past, drug discovery has been a sequential process that involved researchers in the fields of biology and chemistry. Biologists would identify the disease causing agent and then chemists would find all the chemicals that can react with the agent this is a tedious and time consuming process. Drug discovery then involved screening vast randomized chemical libraries against a small number of pharmacologically relevant and in some instances poorly defined biological targets. Subsequently, new knowledge based approach emerged i.e. a more comprehensive, systems biology based approach to identify biological functions, cellular processes and disease mediated processes; this increased the probability of success in drug discovery process [19]. Unearthing the complex physiology of receptor mediated signaling and linking these signaling networks to disease presents exciting opportunities in the identification of multiple targets and creates new and viable options for therapy. Systems biology links all the omic data which would lead to a novel therapy. All the omic data or multi level information reveals insight about pathways and helps in focusing on better therapeutic agents. Systems biology provides a powerful means for validating new drug targets and improving the rate at which pharmaceuticals are identified [20]. Steps to identify effective new drugs:

1. Identify the drug target

2. Identify the drug that will perturb the target
3. Assess the possible side effects and pharmaceutical properties of the drug before its deployment in clinical trials [21] [22]

The researchers of the individualized evidence based medicine or in other words personalized medicine are working towards identifying the bio-markers that can be used to classify the patients into subgroups. Then at a later stage drugs can be designed for each sub-group. The personalized therapy can be a possibility only if the classification can occur and only if the medical information gets integrated. There has been a lot of resistance with the personalized medicine or medicines being made for each individual subgroup. One of the main disadvantages being, groups of patients characterized by less-profitable genotype are at risk of becoming therapeutic “orphans” [16][23].

2.3.1 Impact of Systems Biology in Drug Discovery

Systems biology helps in early diagnosis, disease stratification, individualized therapy (personalized medicine) and ultimately developing preventive drugs depending on both genetic and environmental considerations. Biological networks that interrelate the elements of systems and characterize the flow of information that links the elements and their networks to an emergent biological process. The key idea is to identify the perturbed network. Systems biology can be used for identification, recognition of drugs and their side effects before clinical trials. All these metabolic analysis is done using spectroscopic tools which enumerate thousands of metabolic
products in biological fluids.

Schneider states: “Systems Biology not only will help you identify a disease target to start your drug discovery efforts, but that target will be put in biological context and you will understand the pathway it’s involved in” [24]. This will further help us in:

1. Identification of new uses of the existing targets.
2. Identification of new molecular targets which were not identified previously or not identified with respect to that disease.
3. Deciphering complex signalling relationships

2.4 Digital Biology

Computers are enabling researchers to improve data quality and laboratory efficiency, extend their ability to probe and model complex biological phenomena and enact or adjust to fundamental changes in the conduct of science. As mentioned earlier in this chapter healthcare to a large extent is an information management endeavor. Medical data can be classified into either patient specific or knowledge base. Pharmaceutical and many medical researchers say that the urgent subject of interest is the healthcare integration in other words they want a global database linking all the patients specific information or personal information [16]. They want to develop a digital library for researchers and others.
Three key areas that define the emerging discipline of digital biology are [25]:

1. Scientific Data integration

2. Multi-scale Biological modeling

3. Network Science

2.4.1 Scientific data integration

Biological science has been transformed into an information science. Data sharing is the most important aspect of information science. Challenges faced with integration are:

1. Lack of adequately structured vocabularies and ontologies. Computer readable vocabularies, taxonomies and indexes help in accelerating research

2. Lack of common basis for describing the contents and related data.

3. Absence of common formats or inter-convertible formats that describe the data.

There is a need for integrating or cumulating the healthcare data [26]. The various fragments like prevention, diagnosis, treatment and care need to be integrated in order to leverage personalized medicine. The need for integration gave rise to a new term - *theranostics*(therapy-specific diagnostics) [27]. Systems Biology work mainly deals with modeling, simulation and analysis [28]. The need for developing a standardized language [29]:


1. Each tool uses its own format, these are usually not documented. The result is that a model saved in one tool cannot be loaded into another. This would restrict the exchange of models. This problem can be addressed by introducing a standard format for all tool writers to employ. This standard is called Systems Biology Markup Language (SBML) along with CellML, the introduction of a standard format is beginning to make a significant impact on tool writers, and the majority of the most widely used tools now employ SBML as a means to exchange models. This is called model interchange.

2. The second problem is that many of the tools duplicate each other’s capabilities. Writing simulation tools takes time, and many of the projects are short-lived, which means that the authors are unable to develop the tools very fast. As a result, many of the tools provide similar functionality. Unlike other software development communities, there is little tradition of code reuse in the system biology community. As a result, the community has seen much duplicated effort. The way to resolve this problem is to have a means by which people can reuse an existing code. People developed a software framework called the System Biology Workbench (SBW). The workbench allows different tools to expose programmatic functionality to other tools. This means that a developer can now build on previous work without having to understand in detail the often intricate internal workings of other tools. The workload for the second developer is greatly reduced, and they can instead
concentrate on novel functionality. This work is currently done by DARPA and the DOE.

2.4.1.1 Modeling

Each of the pillars of Systems Biology has a different form of representing the data [28]. The Biologist-cartoon like graphics, Biochemists-chemical reactions and Mathematicians and engineers-mathematical equations.

Systems Biology is a highly interdisciplinary area of research, and the modeling frame work has to be understood by everyone. To overcome all the discrepancies due to the different representations used, the researchers have decided to develop a unified graphical representation language for modeling.

They recently developed a representation which consisted of nodes and arrows. There are two approaches to the systematic representation of molecular interaction-interaction maps or Kohn maps and Systems Biology Graphical Notation (SBGN). It was found that SBGN was a better way of representation because of two reasons-In SBGN different modifications of molecular species are represented by different entities and other reason is that the model in SBGN can be readily converted or read by SBML [30].

2.4.1.2 Simulation and Analysis

An important objective of Systems Biology is to develop predictive models in order to test biological hypotheses and perform in silico tests of new drugs. This
aspect mainly deals with the development of software tools. The main software tools available today are SBW (Systems Biology Work bench) and SBT (System Biology Tool box) [31].

### 2.4.2 Multi-scale Biological modeling

A common feature of complex systems is the presence of a layered or hierarchical topology. Biological information is hierarchical and multi scalar - DNA, RNA, protein, their interactions, networks, cell, organs, individuals, ecologies [32]. Data must be accommodated at various levels of abstraction - molecular, cellular, pathway, organ and even whole organism - and at different stages of analysis. There is a need for computational and theoretical modeling methods that cross scales and that can be validated. "Omic revolution" is a result of this.

### 2.4.3 Network Science

Currently we have pieces and bits of information of a particular patient, diseases, doctor and treatment in different databases. There is an urgent need to integrate all these different databases. Once an integrated database is formed then all the researchers and practitioners can work simultaneously without any delay. The only way that this working can be synchronized with minimum delay is by connecting the real time basic data of the patients to the clinical research group [25].

Linking basic and clinical research through networks will permit scientists to
characterize the function of genetic networks in processes of cellular development, health and disease. These kinds of networks will need transitional middleware. This will ultimately advance evidence-based medicine and improve health or medical care. New requirements are arising even as we work towards a truly effective translational medicine. Specifically, those in clinical settings are finding a need for dealing with multi-scale, complex data and to bring data mining, federation and other modern computer networking strategies to electronic medical record keeping.

In the future, digital biology will be comprised of dense networks of people and resources. Advances in instrumentation, collaboration, data query and analysis offer an array of ways to enhance diagnosis, doctor-patient interaction and other aspects of health care delivery. New grid-based methods allow sharing of geographically dispersed heterogeneous data, software algorithms and computing resources.
Chapter 3

Experimental Setup

3.1 Introduction

This chapter deals with the different steps and methods that are involved in the protein identification process.

3.2 Explanation of the setup on the whole

The experimental procedure encompasses all the steps that are performed on the raw data till the protein ratios are obtained. The whole procedure can be split into the following basic steps - Wet test which consists of preparing the sample and performing the mass-spectrometry; collecting the data and performing data mining (discussed in the 4 and 5 chapters) to obtain the significant proteins. The experimentation part was performed by the research group at Hershey Medical Center. The most common techniques that are usually used for sample preparation
are- Western blotting (used by Keshamouni et al), 2D PAGE (the most commonly used; used in this set up) and 2D DIGE [33]. The first stage consists of preparing the sample and performing the mass-spectrometry, task of the second stage is to obtain peptide sequences present in the sample and the objective of the third stage is to obtain the proteins that these peptide represent.

The experimental setup used in this work is similar to the one used by Keshamouni et al [33]. The main goal or aim of the experiment is identification of the proteins present in the sample. This is a two step process, the first step involves mapping the peptides identified onto the corresponding peptide sequences. The second step is the identification of the proteins with respect to the mapped peptide sequences.

The experiment is a two stage process. The task of the first stage is to obtain peptide sequences. For this MALDI-TOF coupled Tandem mass spectrometry of the iTRAQ labeled samples was used. The task of the second stage is to obtain the proteins that these peptide sequences represent. For this a software was used to identify the protein obtained from the peptide sequence. ProteinPilot was used in our experimentation to identify the proteins whereas Keshamouni et al [33] used Mascot search.

Keshamouni et al classified the proteins depending on their values into up-regulated and down-regulated classes [33]. They also obtained a protein interaction network using Metacore software. To their surprise they found that all the up-
regulated and down-regulated proteins had high degree of connectivity among their corresponding classes. In this thesis work two kinds of analysis are performed to identify the significant proteins that are related to diabetes; they are described in detail in the next two chapters. Chapter 4 discusses the data mining approach and chapter 5 discusses the complex networks approach. In the data mining approach the proteins are classified into three classes (up-regulated, down-regulated and no-change) depending on the protein ratios. In the complex networks approach the sample proteins (Hershey Medical Center protein data) were mapped onto the human protein interaction network (where nodes are proteins and any interaction between them is represented as an edge between them) and performed further analysis to identify the most central and influential nodes (proteins).

### 3.3 Mass Spectrometry

Mass Spectrometry is a technique used to determine the composition and abundance of the atoms in a molecular substance, starting with a very small amount of sample. Mass Spectrometry is a method of identifying and quantifying chemical compounds by converting them into ions and deflecting them using an electric or magnetic field.

The deflection of each ion is proportional to its mass-charge ratio (m/z), and this is used to generate a spectrum which captures the m/z ratio as well as the relative abundance of the impinging ions. Mass spectroscope consists of an ion-
ization source, a mass analyzer (measures the mass to charge ratio) and detector (registers the number of ions at each mass-charge ratio) [34].

The most common protein ionization techniques used to volatize or ionize proteins or peptides are - Electron Spray Ionization (ESI) and Matrix Assisted Laser Desorption (MALDI). ESI ionizes the substrate in the liquid state, therefore readily couples with the liquid separation tools. MALDI sublimates and ionizes the samples out of the dry, crystalline matrix via laser pulses. MALDI is used to analyze simple peptide mixtures, whereas ESI is used to analyze complex samples.

There are four types of mass analyzers used in proteomic research. They are - Ion trap, Time of Flight (TOF), Quadrupole and Fourier Transform ion Cyclotron (FT-MS). In the ion-trap analyser the ion is first captured or trapped for a certain time interval and are then subjected to MS. They are robust, sensitive and relatively less expensive; they have a very low mass accuracy, the reason being the ions are accumulated before mass spectrometry this leads to space charge distortion. Most of the proteomic data published today has been obtained this way. The Fourier Transform ion Cyclotron (FT-MS) captures ions in a vacuum magnetic field. This is very sensitive, provides a high degree of mass accuracy and resolution. The drawbacks of this process is the high cost and complexity. MALDI is usually coupled with Time of Flight (TOF) analyzer that measures the mass of intact peptides. TOF is simple, provides high mass accuracy and resolution; the only drawback of this method is that it needs purified targets [34][35].
<table>
<thead>
<tr>
<th>Mass Analyzer</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ion-Trap</td>
<td>Robust, sensitive and less expensive</td>
<td>Low mass-accuracy</td>
</tr>
<tr>
<td>Fourier Transform Cyclotron</td>
<td>Sensitive, mass accuracy, high resolution and dynamic range</td>
<td>High cost</td>
</tr>
<tr>
<td>Time of Flight</td>
<td>Simple, high mass accuracy, high resolution and sensitivity</td>
<td>Need purified target proteins</td>
</tr>
<tr>
<td>Quadrupole</td>
<td>High mass accuracy</td>
<td>Complicated</td>
</tr>
</tbody>
</table>

Table 3.1. Comparison table for Mass-Analyzers

Figure 3.1. Mass Spectrometer and its components.

3.4 ICAT and iTRAQ

The second generation proteomics emerged as a new paradigm for drug discovery, understanding the biological processes and the development of biotechnological devices[36]. All this started with the process of chemical labelling i.e. ICAT and iTRAQ. Chemical labelling method was mainly used to analyse the differential expression of the substrate. Gygi et al developed the ICAT (Isotope Coded Affinity Tags) methodology [37]. In this method two samples are tagged with chemically
identical substances (differ in their weights) i.e. isotopic tagging [37]. The main disadvantage of this method was that it could only be used to analyze protein which contain cystein [36]. To overcome this disadvantage Applied Biosystems group developed the iTRAQ labeling. This method uses isobaric reagents for tagging. These reagents have the same mass, upon fragmentation produce different reporter ions (m/z= 114,115,116,117) which is used for quantification of four different samples.

### 3.5 Summary of the Experimental Setup

The experimental setup is a three stage process. The experimental setup used by the research group at Hershey Medical Center is similar to that used by Keshamouni et al.

**Step 1 Wet test:** This step consists of preparing the sample and performing the mass-spectrometry.

**Step 2 Data:** This step encompasses all the processes that are involved in obtaining the protein ratios from the mass spectroscopy m/z data.

**Step 3 Data mining:** We performed data mining (discussed in Chapter 5) and network mining (discussed in Chapter 6) on the data obtained.

The steps followed by Keshamouni et al and the research group at Hershey Medical Center were the same. The techniques used were different. The former
used Western Blotting for wet test whereas the later used 2D-PAGE; both of them used MALDI-TOF Tandem mass-spectrometry and applied iTRAQ labeling and for protein identification former used Mascot search and we used Protein Pilot [33].

3.6 Data Description

Dr. Stanley’s group at Hershey Medical Center collected biological samples from patients undergoing gastric bypass. These patients were diabetic before and were to be freed from diabetes after the surgery. We have obtained three datasets from Hershey Medical Center. Each dataset contains data of two patients. The obtained dataset is the output log of the experimentation which consists of - MALDI TOF, quadrupole and iTRAQ. The four iTRAQ reagent labels represent- 114: Patient A before surgery, 115: Patient A after surgery, 116: Patient B before surgery and 117: Patient B after surgery.

The mass spectrometric log is read by the Protein Pilot software to give us the protein ratios. The output obtained from each dataset through Protein Pilot displays three ratio values for each protein. These ratios depend on which denominator is selected (the iTRAQ reagent wrt what we want the relative abundance). Here we have chosen 114 (later 116) as the denominator. The different ratios can be read as follows-

1. When 114 is the denominator: 115/114- Ratio of Patient A after to before surgery, 116/114- Ratio of Patient B before to Patient A before surgery and
117/114- Ratio of Patient B after to Patient A before surgery


For our analysis we have only used the protein ratios obtained by 115/114 and 117/116. This is basically the ratio of patient A after by before and patient B after by before. We consider these ratios because this will help us overcome the problem of each patient is unique and makes further analysis easier as this ratio can be looked as a number rather than a ratio for further analysis like classification.

So, in our case a ratio $r_{ij}$ represents the ratio of protein i in $j^{th}$ patient.

In summary the data we started with was -

1. On an average 200 unique protein ratios per patient,

2. Protein ratios from six unique patients, and

3. All the patients were diabetic before surgery and were freed from diabetes after the surgery. All the patients underwent a change of state with respect to surgery.
4.1 Introduction

In the past, genomics and proteomics study was focused on one gene or one protein at a time. With the advent of high throughput technologies the activity and interaction of thousands of genes or proteins can now be captured and measured simultaneously. The nature of research has transformed from a hypothesis driven one to a data driven one; also revolutionizing the study of complex human diseases. All these changes will give rise to a patient tailored therapy or in other words lay a path for a personalized medicine phase.
4.2 Literature Review

4.2.1 Genomic and Proteomic data

Genomics is the systematic study of genes, their functions, and their interactions. Proteomics is the study of proteins, protein complexes, their localization, their interactions, and their posttranslational modifications. Proteins are made of amino acids linked by peptide bonds to form polypeptide chains. A typical protein contains 200 to 300 amino acids. All amino acids have a central carbon atom which makes four attachments. Three of the attachments are the same for all the different amino acids (a single hydrogen atom, a COOH group on one side and an NH$_2$ group). The fourth attachment differs for each of the 20 amino acids. Proteins take many physical forms depending on the function they perform. Composition wise genomes and proteomes are entirely different in the sense that genes are made up of four basic nucleic acids (A, G, T, C), whereas the basic sub units of proteomes are twenty amino acids.

The gene data is obtained from the DNA (Deoxy Ribose Nucleic Acid) microarray data and the protein data is obtained from the peptide mass spectrometric data. Gene and protein data is not obtained directly, they are usually inferred from their components (gene - DNA and protein - peptides). The data obtained in both the cases is a ratio. Because of these reasons one can use the same algorithms on the genomic and proteomic data in the sense that the same kind of methodology
can be applied to identify the significant genes and the significant proteins from the corresponding genomic and proteomic data.

In recent years there has been a shift in the trend from genomics to proteomics. The number of human genes detected was much below the number of genes predicted [1]. The 30,000 human genes reported are incapable of expressing the necessary number of proteins. The number of genes detected is well below that predicted for human, indicates that humans have increased their complexity by using the preliminary building blocks in diverse ways. This led to the examining of the products encoded by the genes i.e. the proteins. This made a better correlation with respect to a disease. This is the study of proteomics in the manner similar to genomics. The complexity of analysis also increased from the four building blocks of genomes to the study of 20 individual amino acids that makeup the protein. This led to the change in the research focus from genes to proteins.

4.2.2 Biomedical Datamining

High throughput technologies have made an undoubted impact in the fields of bio-medical research, healthcare and pharmacology. Researchers working in this field feel that there is a long way to go from bench to bedside. High throughput technologies generate terabytes of genomic and proteomic data which provide the researchers with unprecedented analytical and computational challenges. There are two main reasons for this[38]:
1. Ethical issues, cost and time involved in running these experiments.

2. Most life science studies include a modest number of cases, \( n \) (\( n \) lies within the range of few dozens to few hundred); and a large number of variables, \( p \) (\( p \) lies in the range of tens of thousands). This problem is called the *curse of dimensionality*.

The current situation is that there is a large amount of data with no appropriate tools. Most bio-medical researchers are working towards developing appropriate or suitable data mining tools.

Barry and Linoff define data mining as "the exploration and analysis, by automatic or semi automatic means, of large quantities of data in order to discover meaningful patterns and rules" [39]. The exploratory approach is usually used to obtain a basic understanding of the different qualitative and quantitative (data visualization, clustering, data reduction) aspects of a given set. These techniques are generally used to generate hypothesis. Analytical techniques are normally concerned with more precisely formulated questions or testing a hypothesis. The commonly used analytical tools are-data classification, correlation and sensitivity analysis. Pattern recognition and data mining techniques are often used when the objective is not clearly known. The most common questions that researchers try to find answers for in the area of genomics and proteomics are-

1. Are there any interesting patterns in the data set
2. Are the array profiles characteristic of a particular class

3. What features (eg: genes, proteins) are most important

A pattern can refer to a group of data or in other words can be called a cluster. Clustering (the process of forming clusters) is an unsupervised learning method; as the process is not guided by any pre-selected class labels but by similarity or dissimilarity. Clustering refers to an exploratory approach to reveal relationship that may exist in the data. Clustering methods attempt to maximize the similarity between elements of a cluster, while minimizing the similarity between different groups.

Given n cases, let k be the number of predefined class labels-

\[ c = \{c_1, c_2, ..., c_k\} \]

class labels are arbitrarily relabeled as \( Y = 1, 2, ..., k \).

Each case \( x_i \) is described by p-observations, \( x_i = \{x_1, x_2, ..., x_p\} \). Each class has exactly one class label i.e. \( \{x_i, y_i\} \subset X \).

Classification is a process of learning from examples, in which the objective is to classify an object into one of the k-classes, i.e. to predict \( y_i \) from \( x_i \). Simon in his paper mentioned that the most common errors of micro-array data mining are the use of clustering techniques for a classification task [38]. No classifier is inherently superior to any other. This is called the No free lunch theorem. Somorjai criticized the sophisticated methods used to classify the microarray data; he felt that simple algorithms work as well as the sophisticated ones [40]. Hastie and Dudoit [41] in their paper proved that simplest method like nearest neighbor classifier often
performed as well as more sophisticated Support Vector Machines (SVMs) method. Most Support Vector Machines perform well on high dimensional data involving only two classes.

### 4.2.3 Data mining technique - Gene Filtering

The identification of significant genes or proteins from the large amount of high throughput data can in other words be called - filtering out (the genes or proteins we do not need) or filtering in (onto the genes or proteins that we feel might have a significance). The researchers gave this process the name *Gene Filtering* [42].

Composition wise genomic and proteomic data are entirely different in the sense that genes are made up of four building blocks i.e. nuclei acids \((A, G, T, C)\). Whereas, the proteins are made up of twenty individual component amino acids. Though they are entirely different composition wise one can draw a few parallels between the genomic and proteomic data. The gene data is obtained from the DNA (Deoxy Ribose Nucleic Acid) microarray data and the protein data is obtained from the peptides mass spectrometric data. The obtained data in both these cases is a ratio. Because of these reasons one can use the same algorithms on the genomic and proteomic data in the sense that the same kind of methodology can be applied to identify the significant genes and the significant proteins from the corresponding genomic and proteomic data. As already discussed in the previous section about the similarities of the genomic and proteomic data. We have
applied the same gene filtering technique concept to the protein data to identify the significant proteins.

The filtering process is a two step process. The first step deals with the identification of genes or proteins of interest and the second stage deals with ranking of these significant genes or proteins in the order of their significance [43]. This filtering technique was used on genes and thus called gene filtering and the steps involved are mentioned below-

**Step 1 Gene Screening**

**Step 2 Gene Ranking**

Gene filtering is a technique used to filter genes based on the variance in their expression levels i.e. identification of genes that are differently expressed.

**4.2.3.1 Techniques used for screening**

The problem of gene or protein screening can be reformulated as a problem of significant gene or protein identification. Data mining and statistical techniques are the most commonly used screening techniques. The screening techniques usually use false discovery rate (FDR), variance-normalized differential expression levels and minimal acceptable difference (MAD) as the criteria [44] to segment genes or proteins. A threshold value for each of these criteria and only those genes whose values are above the threshold for each of these criteria are selected. It was shown
that by screening the data according to various criteria (more than one) thresholds, a significant improvement in results was obtained.

The next question or the need after identifying the significant genes or proteins would be to know which of these significant genes are the most important. For example, if one is researching genes or proteins in order to identify the drug targets, the ranking of the genes or proteins would identify the order for clinical trials.

4.2.3.2 Techniques used for Ranking

Ranking is the process of ordering the proteins with respect to their level of significance. Hero et al [44], proposed multi-criteria methodology to rank the genes identified in the gene screening stage. They proposed a pareto optimal approach (non-dominance) in which the probability of genes lying on the pareto front was used as a ranking criteria. Hero and Fleury [44] were dealing with a large data set and were concerned with alleviating sampling error. To this effect they proposed a method called Resistant Pareto Front (RPF) approach [43]. In this paper they proposed a method called leave-one-out. By implementing this method one would obtain a number of sample datasets from the original population. This method is totally data driven and free from any distribution assumptions. In the paper Hero and Fleury introduced a new parameter called relative frequency score for quantification. This parameter depends on the number of times a particular gene occurs in the pareto front after the whole, leave one out occurs.
In my analysis we will be using the same procedure to obtain the filtered protein data i.e. the significant proteins and also to order the proteins with respect to their level of significance. In order to perform such analysis we wish to identify criteria which can be used to identify the relevant proteins.

In this work we used the protein ratio, frequency (the number of patients in which a protein occurs) and the occurrence (the number of times a particular protein occurs in the database that we obtained from the Resistant Pareto Analysis). All these terms have been discussed in detail in the section 4.2.3. We will be using the above mentioned RPF procedure for protein filtering but with a little variation. We will not be defining a threshold on these values, we will rather classify them into three classes. We will be discussing this procedure in detail in the next section. After we classify these proteins, for ranking we will be using the Pareto front approach. In my analysis I have not used the Resistant Pareto Front (RPF) as method for cross validation. I am using it to define one more criterion for protein screening (occurrence).

The procedure for generating the resistant pareto front is as follows - first six (the number of patients) different data sets are formed with each data set containing the data of five patients i.e. data of one patient is removed from each data set. This is called the leave-one-out approach. Applying the leave-one-out approach we obtained six data sets, in each one patient’s data has totally been left out. Quantification parameter, occurrence has been defined as the number of times the
particular protein has occurred across all the 6 data sets.

4.3 Method

A methodology similar to filtering has been used. The process consisted of two steps. The aim of the first step is identification of the significant proteins i.e. in other words called screening (as discussed in the previous section). The second step is ranking of the proteins in order of their significance; this ranking method has been used to generate one more parameters for screening. The proteins were classified into one of the three classes depending on the protein ratio. The three classes have been divided depending on the range into which the mean of a protein ratio across all the patients ($\bar{r}_i$) is present. The proteins that belong to class 1 are up regulated i.e. the protein value was higher after the surgery compared to before the surgery. The proteins that belong to class 2 are down regulated i.e. the protein value was lower after the surgery compared to the value before surgery. The proteins that belong to class 3 are the ones whose values haven’t changed w r t the surgery.

$$\bar{r}_i = \frac{r_{ij}}{\eta_i}$$

Where,

i-Represents the protein

j- Represents the patient

$\eta_i$-Number of patients in which protein i is present, frequency
\( \bar{r}_i \)-Average ratio of a protein across all the patients

\( r_{ij} \)-the value of the ratio of a protein in patient \( j \)

4.3.1 Class Definition

As already discussed in the previous section the proteins have been classified into three classes w.r.t the average protein values. Proteins that belong to class 1 are all the up regulated proteins. The proteins that belong to the class 2 are the down regulated proteins. The proteins that belong to class 3 remain unchanged. The class definitions are given below:

Class 1:  
\[ C_1 = P_i / ((\bar{r}_i > 1 + \Delta_i) \cap (\eta_i > 1) \cup ((\bar{r}_i > 1) \cap \eta_i = 1) \]

Class 2:  
\[ C_2 = P_i / ((\bar{r}_i < 1 + \delta_i) \cap (\eta_i > 1) \cup ((\bar{r}_i < 1) \cap \eta_i = 1) \]

Class 3:  
\[ C_3 = P_i / ((1 - \delta_i) \leq \bar{r}_i \leq (1 + \Delta_i)) \cap (\eta_i \geq 1) \]

Where,

\[ \Delta_i = 2\sigma_i \]

\[ \delta_i = \frac{2\sigma_i}{1+2\sigma_i} \]

\( \sigma_i \)- Standard Deviation of Protein i across all the patients

4.3.1.1 Results Classes

There are 595 unique proteins. Class 1 consists of 53 proteins, Class 2 contains 82 and the rest i.e. 460 proteins belong to Class 3. The proteins that underwent
change are the proteins of our interest or in other words *significant proteins*. Here the proteins of Class 1 and Class 2 are the proteins that underwent a change (protein ratios have changes w.r.t the surgery) and are the proteins of interest. The proteins of Class 1 are the up-regulated protein and the proteins of Class 2 are the down-regulated proteins. The dimensionality of the proteins has decreased to 135 proteins.

![Figure 4.1](image.png)

**Figure 4.1.** Plot of Class 1 Proteins - Average Ratio and Frequency.

From the above three graphs it is evident that most of the proteins have a frequency (the number of patients in which the protein occurs) of 2, 4 or 6. Proteins of class 1 and class 2 have a frequency of 2 or 4. The average protein ratios lie in the range of (0,2] for most of the proteins. As already discussed frequency is an attribute of a protein; frequency is defined as the number of patients in which the protein occurs.
4.3.2 Level of Significance

The next step after the identification of significant proteins i.e. the proteins that underwent a change with respect to the surgery (class 1 and class 2 proteins). There is a need to identify one more parameter that can help decrease the dimensionality of the significant proteins. As already discussed in the section 4.2.3 the level of
significance of the proteins can be measured in terms of the average ratio, frequency and occurrence. The various methods of ranking relevant to this kind of data has already been discussed in the ranking section 4.2.3.2. The Resistant Pareto Front method of approach was the one used for obtaining the new required parameter called occurrence.

![Figure 4.4. Proteins - Resistant Pareto Front.](image)

4.4 Results

The main aim of this thesis is to identify the proteins that have affected the change of state of the patient. Here the change of state of the patient is disappearance of diabetes. The proteins of our interest are-

1. The proteins that underwent a change, in other words are the proteins that
<table>
<thead>
<tr>
<th>Accession Number</th>
<th>Accession Number</th>
<th>Average Ratio</th>
<th>Class</th>
<th>Frequency</th>
<th>Occurrence</th>
</tr>
</thead>
<tbody>
<tr>
<td>P00450</td>
<td>CERU − HUMAN</td>
<td>1.0724759</td>
<td>1</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>P01008</td>
<td>ANT3 − HUMAN</td>
<td>1.211936321</td>
<td>1</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>P02760</td>
<td>AMBP − HUMAN</td>
<td>1.125452234</td>
<td>1</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>P09871</td>
<td>C1S − HUMAN</td>
<td>1.138375957</td>
<td>1</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>Q06033</td>
<td>ITIH3 − HUMAN</td>
<td>1.446693682</td>
<td>1</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>Q14520</td>
<td>HABP2 − HUMAN</td>
<td>1.154867319</td>
<td>1</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>Q96PD5</td>
<td>PGRP2 − HUMAN</td>
<td>1.394025962</td>
<td>1</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>O14791</td>
<td>APOL1 − HUMAN</td>
<td>0.807064182</td>
<td>2</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>P02743</td>
<td>SAMP − HUMAN</td>
<td>0.698867439</td>
<td>2</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>P05156</td>
<td>CFAI − HUMAN</td>
<td>0.908310708</td>
<td>2</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>P36955</td>
<td>PEDF − HUMAN</td>
<td>0.887606398</td>
<td>2</td>
<td>4</td>
<td>6</td>
</tr>
</tbody>
</table>

**Table 4.1.** Proteins of interest

1. Proteins of interest belong to Class 1 or Class 2.

2. The proteins that have highest occurrence of six, and

3. The proteins that have highest frequency.

The results indicate that there are eleven proteins of interest or in other words there are eleven proteins that satisfy all the three required criterion. The obtained eleven proteins and their details are tabulated in table 4.1.
Chapter 5

Network Mining

5.1 Introduction

The aim of this chapter is to rank the test proteins (obtained from Hershey) based on their role in the human protein interaction network. A human Protein-Protein Interaction network was built using the data provided by College of Life Sciences at the University of Dundee [45]. The Dundee dataset was used as it provided a global picture of the human PPI network. The ranking is mainly based on the "centrality" of the protein in the network. Other significant structural patterns like the scale free connectivity, existence of a hierarchy and modular nature (existence of clusters) in the protein network were observed. Interestingly, most of the proteins seem to lie in a single cluster/module. More structural analysis needs to be performed to identify the role of a protein in diabetes based on its position in the protein interaction network. The proteins in the University of Dundee dataset
followed the International Protein Index format (IPI). The protein dataset from Hershey Medical Center followed the UniProt format. Therefore, protein names needed to be converted from UniProt format to IPI format. After standardizing the data, human protein interaction network was built (see Fig: 5.1).

In order to obtain deeper understanding of the topological properties of the network, a number of statistical tests were performed. The final step was ranking of proteins and identifying any significant patterns.

5.2 Literature Review on Complex Networks and Systems Biology

There are a lot of complex systems around us; climate, cells, nervous system and all the living things. These complex systems are very difficult to understand. As a first attempt to understand these complex systems researchers used the reductionism approach. This is an approach of understanding the entire system based on the description of its sub-systems and ignoring the relation between them [46]. This was a very myopic approach because most complex systems could not be understood by the functioning of their components. All the complex systems exhibited emergent properties i.e. whole system is not equal to sum of its components. In this review I will discuss about the reductionistic approach, holistic approach which is by far the best approach, systems biology (integrative approach towards under-
standing biological systems) and complex networks (which is by far the best tool for modeling these complex systems).

The reductionistic approach as already defined is a very successful approach in identifying the components and their functionality. This approach was very successful in making advances in the fields of physics and chemistry. This was not a very successful approach when it comes to biology, as the components or sub-units cannot be successfully isolated. Currently in biology we are at a stage where the building blocks have been identified and the integration process still remains. This integrative attempt developed as a new inter-disciplinary area of study called Systems Biology. It is evident that the isolated knowledge is not good enough to understand the biological systems. For example, $P53$ protein was first identified as an oncogene and after ten years of research scientists found that it was a tumor suppressor [47][48].

A disease or a biological function can only rarely be attributed to a molecule - it is usually an interaction of numerous molecules that perform a function or bring about a change. So to understand or even to identify the critical component one needs to look at the entire system [1]. This is a holistic approach and is called the Systems Biology based approach. Systems biology is an approach that combines the old hypothesis driven approach and the new discovery science (data driven approach) approach [49]. The main aim of the Systems Biology approach is to collect the data from different levels of the biological systems and then to
integrate them in a way so as to generate a predictive model [50]. System Biology gained a great deal of popularity because of the advent of drug discovery especially drug target (is the biological component which when acted upon by a chemical molecule helps in reducing or eliminating the abnormal functioning of the system) identification and personalized medicine (based on the principle that people are unique and so should their therapy be) [51]. Molecular Biology reveals a great deal of multitude of facts but this by itself is not sufficient to understand biological systems.

The main reason for adopting a systems biology type approach is not only to model the enormous number of interactions between various components of the human body but also capture features like robustness and modularity. A natural way to represent such discrete interacting units is a graph or a network. The nodes represent the biological components (eg: proteins, genes, metabolites) and the interaction is represented by an edge (eg. protein interactions, gene regulation, metabolite reactions) [52].

5.2.1 Complex Networks

A network is represented as a graph $G = \{V, E\}$ where $V = \{v_1, v_2, ..., v_N\}$ represents the set of vertices and $E = \{e_1, e_2, ..., e_M\}$ represents the edges. A pair of nodes is connected by an edge and if the edge is directed then the graph is called a directed graph and is represented by $G = \{V, \overrightarrow{E}\}$ and $\overrightarrow{E}$ represents an ordered
set. The network can also be represented as an adjacency matrix $A_{N \times N}$. The element $a_{ij} = 1$ if the nodes i and j are connected by an edge and is 0 otherwise. The matrix is symmetric in an undirected graph as $a_{ij} = a_{ji}$. This does not hold in the case of directed graphs.

The current natural or real networks are made up of a large number of nodes and edges. These networks cannot be understood through visual inspection because of their size. The complex network literature has many topological measures that can be used to understand the network. We will discuss the topological measures in the coming sections.

5.2.1.1 Node Attributes

A node or vertex in a network can be a person, cell, gene, protein, web page or even a power generators. In this section we will talk about the individual node properties which will later make it easier to understand the whole network properties and topologies. The node attributes are described below-

- **Degree:**

  Degree is defined as the number of edges connected to a vertex. The degree ($k_i$) of a node $v_i$ is the number of vertices that are connected to the node through edge. This is also referred to as connectivity. From the adjacency matrix $k_i = \sum_j^N a_{ij}$. For directed graphs we distinguish between the in degree $k_{in}$ (the edges that are directed towards the node) and out degree $k_{out}$ (the
edges that are directed outwards from the node). The nodes with high degree are referred to as hubs. A node has a degree and the entire graph follows a degree distribution.

- **Clustering Coefficient:**

Cluster Coefficient is a local measure of quantifying the likelihood of neighbors of \( v_i \) are connected with each other. In sociology literature it is referred to as fraction of translative triples[53]. It is calculated by dividing the number of neighbors of \( V_i \) that are actually connected among themselves by all the possible self links [54]. This is mathematically computed as follows:

\[
C_i = \frac{\text{number of triangles connected to node } v_i}{\text{number of triples centered on node } v_i} \tag{5.1}
\]

\[
C_i = \frac{2e_i}{k_i(k_i - 1)} \tag{5.2}
\]

- **Betweenness Centrality:**

The betweenness centrality \( b_m \) of a vertex \( v_m \) is defined as the number of geodesic paths between other vertices that run through node \( v_m \) [55][56][57]. Betweenness Centrality is a measure of network resilience-312]it tells us how many geodesic paths that will get longer when the vertex is removed from the network [58].
\[ b_m = \sum_{i \neq j} \frac{\Gamma_m(i, j)}{\Gamma(i, j)} \]  

Where, \( \Gamma(i, j) \) is all the shortest paths between node i and node j; \( \Gamma_m(i, j) \) represents all the paths that pass through node m. The betweenness centrality is a very important measure of the network compared to degree [59], as nodes with high betweenness centrality necessarily need not have a high degree, but play a very important role in keeping the network connected.

- **Shortest Path:**

  A path is the string of nodes relating \( V_i \) and \( V_j \). The *shortest path* connecting \( V_i \) and \( V_j \) is the one where the lowest number of nodes are involved to connect them.

All these topological node attributes are very important and play a major role in understanding the network attributes. For instance, high \( K_i \) for a node might relate to a relevant role, since many other nodes interact with it and a high \( b_i \) implies that a number of nodes are efficiently connected through it.

**5.2.1.2 Network Attributes**

Network parameters help in providing us with an incite into the network with millions of nodes and edges which are otherwise very difficult to understand and analyze.
• Small World:

The real world networks are very large and consist of millions of nodes. Despite their size most networks have a relatively short path length between two nodes. The distance between two nodes is defined as the number of edges along the shortest path connecting them. A most popular experiment was conducted by a sociologist Stanley Miligram, who concluded that there was a path of acquaintance with a typical length of six between most pairs of people in the US. This property is a characteristic of complex networks. For example, movie actor network-the actors are 3 degrees apart and the chemicals in a cell are 3 degrees apart.

• Degree Distribution:

The spread of the number of edges a node has or a node degree of a network is characterized by a distribution function $P(k)$. The average degree of the network is defined as $\langle k \rangle = \frac{2E}{N}$. This $P(k)$ is defined as the probability that a randomly selected node has a degree $k$ or the fraction of nodes in the network that have $k$ degree. The degree distribution of a random graph follows a poisson distribution (a binomial distribution with a large $N$ and small $p$). Researchers found that the degree distribution of most real world networks was right skewed with a long tail. The tail was found to follow a power law degree distribution $P(k) \sim k^{-\alpha}$ or exponential distribution $P(k) \sim e^{-k}$. 
• **Betweenness Centrality Distribution:**

The betweenness distribution is represented $b_i$ vs $k$. There is no dependency between $b_i$ and $k$ in the case of random graphs. In the case of real-world networks like the internet, it is found to follow a power-law distribution.

• **Clustering Distribution:**

The Clustering distribution is defined $C_i$ vs $k$. In the case of random as well as scale-free networks, the clustering coefficient and the degree have no bearing. Contrary in the case of hierarchical networks, the clustering coefficient varies inversely with the degree $C_i \sim k^{-1}$. These types of networks exhibit modularity. A module can be defined as a group of nodes with large numbers of edges falling within the group as opposed to leaving the group. There are several algorithms available to discover the modules in a network.

• **Assortative Mixing:**

Assortative mixing ($r$) is a measure that weighs the correlation among degrees in a graph [38]. In other words, assortative mixing means the preference of high-degree nodes to attach to other high-degree nodes and low-degree nodes to low-degree nodes. Disassortative mixing means that high-degree nodes attach to low-degree nodes and vice-versa. This is the Pearson correlation coefficient of the degrees at the two ends of an edge; value ranges from -1 to 1 [60]. The below-mentioned formula is used to calculate the
assortative mixing of the network for practical purposes. Negative indicates disassortative mixing and positive assortative mixing. The assorative mixing of the protein interaction network was found to be -0.07. This implies that the network is dissortative in nature.

To obtain better or further understanding of the network. The following parameters may also be computed-

\section*{5.2.2 Classes of Networks}

\subsection*{5.2.2.1 Random Graph}

The first serious attempt to generate a random network was by Rapoport and his collaborators [61], this was independently re-discovered by Erdos and Rnyi a decade later [62][63]. They performed extensive analysis and called it a random graph. The first definition for a random graph is that there are $N$ labeled nodes that are connected by $n$ edges, that are selected at random from $\frac{N(N-1)}{2}$ possible edges. In total there are $C_n^{N(N-1)}$ graphs with $N$ nodes and $n$ edges.

A random graph can also be defined as a network that starts with $N$ disconnected nodes and an edge is formed between a pair of nodes with a fixed probability of $p$. In mathematical literature construction of a random network is referred to as \textit{evolution}; the network starts with $N$ disconnected nodes and the graph is formed by successively adding edges randomly. The goal of a random graph is to determine at what connection probability $p$ i.e. the critical property $p_c$, the network
property would arise. Erdős and Rényi assume that every graph has a unique property \(Q\) (e.g.: the whole network is connected) which needs to be maximized. The transition from a property being very unlikely to very likely is swift and happens suddenly.

The **degree distribution** of a random graph follows a Binomial distribution (where pairs of nodes are connected with a probability \(p\) and not with a probability of \(1-p\)) as \(N\) is large and \(p\) is small it can be estimated to be a Poisson distribution.

The **diameter** of the graph is the maximum distance between any pair of nodes. This is calculated for the largest connected component (the complete graph without the isolated nodes). Diameters of random graphs is found to be small and proportional to \(\frac{\ln(N)}{\ln\langle k \rangle}\). A general conclusion is that for most values of \(p\), almost all graphs have precisely the same diameter.

The **clustering coefficient** of a complex network is large. In the case of a random graph the probability that two neighboring nodes of a node are connected is equal to the probability that two randomly selected nodes are connected. Thus, \(C_{\text{rand}} = p = \frac{\langle k \rangle}{N}\). From this we feel that a log log plot of \(\frac{C_{\text{rand}}}{\langle k \rangle}\) vs \(N\) is a straight line with a slope of -1. But the truth is that clustering coefficient has not bearing on the size of the network.
5.2.2.2 Small Work Network

Small world networks have a large clustering coefficient which is independent of the network size $N$ but depends only on the co-ordination number (this is property of a lattice). A small world network starts with a fully regular network with $N$ nodes in which each vertex is connected to $k$ nearest neighbors. Then each edge is randomly rewired with a probability of $p$. These networks were developed to explain social networks (nodes are people and edges represent a relationship between them); a person starts with a certain number of friends who are diametrically close to each other and then their friends change with respect to time and they are no longer their neighbors.

Watts and Strogatz proposed a one dimensional model that was between a regular lattice and a random graph [64]. The algorithm behind this model is -

1. Starts with a ring lattice of $N$ nodes each node is connected to $\kappa$ of its nearest neighbors ($\kappa/2$ nodes on each side).

2. The edges of the network are rewired excluding self-link and duplication with a probability of $p$. This accounts for the randomness of the system.

When the rewiring probability is 0 the network has a high clustering coefficient but a large average shortest path length, as $p$ tend to 1 the clustering coefficient degreases and the average shortest path length decreases. As the rewiring probability increases from 0 to 1 the network changes from a lattice to a random graph. This model is between an ordered finite dimensional lattice and a random graph.
\[ \ell(0) = \frac{N}{2k} \gg 1 \text{ and } C(0) = \frac{3}{4} \quad (5.4) \]

\[ \ell(1) \sim \frac{\ln(N)}{\ln(k)} \text{ and } C(1) \sim k/N \quad (5.5) \]

The first set of equations show that the systems is an ordered lattice. It can be seen that the average path length scales linearly with system size and the clustering coefficient is large. The second set of equations show that the model converges to a random graph. It is seen that the average path length scales logarithmically with system size and the clustering coefficient decreases with N.

The \textbf{average path length} changes as a function of rewiring probability p. When p is small \( \ell \) scales linearly with the system size N; when p is large \( \ell \) scales logarithmically with the network size N. The rapid drop in the path length is because of the appearance of shortcuts. Every shortcut connects widely separate parts of the network and thus decreases the average path length of the network. \( \ell \) does not begin to decrease until \( p \geq \frac{2}{Nk} \).

The \textbf{clustering coefficient} displays duality for a wide range of values of rewiring probabilities. In the regular lattice configuration (p=0) the clustering coefficient does not depend on the network size but only the topology. When the edges are rewired the clustering coefficient remains close to C(0) up to a relatively high p.
In the WS model at $p=0$ each node has the same degree $k$. A non-zero $p$ broadens the range of degree keeping the average degree at $\kappa$. The degree distribution is similar to the random graph and peaks at $\langle k \rangle = \kappa$ and decays exponentially for large $k$.

### 5.2.2.3 Scale Free Network

The networks discussed till now model networks using the observed real world properties like the degree distribution or transitivity, these networks do not help in understanding the reason why these properties exist. Scale free networks capture the evolution and dynamics of the system. This class of models are the most used and studied models [65] [66][67][68]. Derek de Sotta Price around 1965 studied the scientific paper citation network [69] and found that the in-degree and the out-degree follow a power law distribution. Price’s work was based on Herbert Simon [70] who said power-law arises when ”the rich get richer” (amount you get goes up by the amount you already have). Price called it *cumulative advantage*. It is referred to as *preferential attachment*, the name coined by Barabasi and Albert [71].

Goal of random graph and WS model is to construct a graph with correct topological features. The scale free model is used when the emphasis is on capturing network dynamics, topology being a side product. The most popular and widely used scale free model is the Barabasi and Albert model[72]. Random graphs and
WS model the number of nodes in the network remains constant. But, in real world systems are open and they *grow* i.e. the number of nodes keep increasing. For example, the world wide web system, new web pages are created everyday and the number of nodes grow at an exponential rate. In the random graph and WS model two nodes are connected independent of the node degree i.e. new edges are placed randomly. Most real world networks exhibit *preferential attachment*. Barabasi and Albert in their model took these two features (growth and preferential attachment) into account [73]. Their algorithm is as follows:

**Growth:** Starts with a certain number of nodes and at every time step a certain number of nodes are added and;

** Preferential Attachment:** A new node that enters the system connects to node $i$ with a probability $\Pi$ which is a function of degree and is given by: $\Pi(k_i) = \frac{k_i}{\sum_j k_j}$.

For a network to be scale free it needs to have both the conditions - growth and preferential attachment.

The *average path length* ($\ell$) of these networks are smaller than that of a random graph of same order. This implies that the heterogeneous nature of scale free networks gets the nodes closer compared to a homogeneous network topology. The *clustering coefficient* is also higher for these networks compared to a random graph of same order. The clustering coefficient of a scale free network increases as the network size increases. The *degree distribution* of scale free networks follow a power law distribution; $P(k) \sim k^{-\gamma}$ where $\gamma$ lies between 2 and 3. This implies that the there are few high degree nodes and a large number of low
degree nodes. This kind of heterogeneous networks are robust to random attacks and break fall apart in the case of targeted attack. The between centrality distribution of this network follows a power law distribution. There are a lot of models that were developed basing on the Barabasi and Albert model. The new models are developed by with more focus on either growth or preferential attachment [74][75][76].

5.2.3 Examples of Networks

All the networks of the real world belong to one of the following four categories namely-social networks, information networks, technological networks and biological networks. In this review I will discuss more about the biological networks (the focus)[72].

5.2.3.1 Social Networks

The first and longest studied networks are in the social sciences area. In social networks nodes are people and any interaction between them is represented with an edge. Social networks are used to model friendship networks, business relations between companies and intermarriages in families [72]. The most famous works are of Jacob Marenco - model friendship patterns in small groups, Davis and collaborators studied the social circles of women in a particular US city, Elton Mayo studied the social circle of Chicago factory workers and Miligrams experi-
ment of six degrees of separation. Traditional methods of studying social networks were inaccurate, subjective, small in sample size and labor intensive. To overcome these problems researchers started to study - collaborator (movie actor, director, scientific paper collaboration network) and personal communication networks (telephone, email, letters) [73].

The average path length of the movie actor collaboration network is close to that of a random graph with the same size and the clustering coefficient is much much higher than a random graph . The degree distribution of the movie actor network has a power-law tail for large k, \( P(k) \sim k^{-\gamma_{actor}} \) \( -\gamma_{actor} = 2.3 \pm 0.1 \) [77]. A scientific collaboration network can be constructed similar to that of the movie actors network, the nodes in this case are the scientists and two nodes are connected if the two scientists have written an article together. These networks show small average path length and high clustering coefficient. The degree distribution is a perfect power-law with an exponential cutoff [78].

5.2.3.2 Information Networks

The second class of networks are the information networks or the knowledge networks. The most common networks that fall into this category are the scientific citation network, scientific patent networks and the world wide web. These are called information networks as each node contains information and we are trying to make a link between these sources of information. In the case of a scientific
citation network, the nodes are papers and if a paper B cites a paper A there is an edge going from node B to node A. These are acyclic networks (a paper can cite a paper that is already published)[79]. The probability that a paper is cited k times (in degree) follows a power-law but the out degree has an exponential tail. A World Wide Web can be modeled as a network with web pages as nodes and a hyperlink between them is represented by an edge. This is a cyclic network with a power law degree distribution. Despite the large number of nodes, the WWW displays the small world property. The average path length of the WWW inspite of the large number of nodes is much lesser than a comparable random graph. The hyperlinks are directed, the clustering coefficient considering the edges are bidirectional is much higher than a random graph [73].

5.2.3.3 Technological Networks

The technological networks are the man made networks. These were made for distribution of commodity or resources. The first studied technological network is an electrical power grid network. In these networks nodes are generators, transformers and substations, and the edges are high-voltage transmission lines. The average path length was equal to a random graph of comparable size. The clustering coefficient was much higher than a random graph. The degree distribution was found to be exponential [74] [73]. The internet popularly studied network of this class. These networks are studied at two levels-the inter domain level [80](nodes
are domains which can contain a number of routers and edges are the inter domain links) and the router level [81] (nodes are routers and any physical connection between them is represented with an edge). The degree distribution of the inter domain as well as the router follows a power law. The average path length of the inter domain networks and the router is much smaller than a comparable random graph. The clustering coefficient is higher in both the cases as compared to a random graph.

5.2.3.4 Biological Networks

The reductionistic approach of molecular biology has helped in providing knowledge of individual cellular components and their functioning. There are three key features which are common to most of the complex biological systems (a) Emergence: highly complex and non-trivial behavior emerges as a result of interactions between simple system components (b) Robustness: the biological system exhibits remarkable robustness to changes in external environment (c) Modularity: the system can be decomposed into groups of components or modules each of which performs a particular task or function [82]. If the clustering coefficient of the system varies inversely with the degree \( Cαk^{-1} \), the network is said to exhibit modularity. All these features have been widely studied in the complex networks (e.g. www, Protein interaction networks) research and thus can be used in systems biology research. This modeling technique, also referred to as ”complex networks”; has
been immensely successful in uncovering structural and functional organizations in several biological processes. In this section we will discuss about the critical networks that regulate or control cellular activities. They are - transcription regulatory network, protein-protein interaction network and the metabolism network [52] [73] [72].

**Transcription Regulatory networks** are the systems that regulate the gene expression in the cell. These networks capture the relation and functioning of the action between the Transcription Factor (TF) and the Transcription Gene (TG). The TF responds to changes in the cellular environment by regulating the transcription of the TG. These systems very complex not because of the number of different genes but because of the elaborate regulation mechanism [83]. The TF and TG are represented as nodes in the network (see figure A.1). The TF are always the source nodes and the TG the sink nodes. The relation between them is represented by a directed edge. The out-degree distribution of the TF follows a power-law whereas the in-degree follows an exponential distribution. This implies that most TF regulates few TGs but a few TFs interact with many TGs, but most TGs are regulated by the same TF. These networks are modular in nature and are made of motifs. The most common motifs found in TRN are - single input, multiple input and feed forward loops [84].

The study of **Protein Interaction Network** is important for various biological processes such as cell to cell communication, the perception of environmental
changes and protein transportation and modification. In a protein interaction network the nodes are proteins and any physical interaction between them is represented by an edge (see figure A.2). The protein interaction of yeast S. cerevisiae [85], bacterium Helicobacter pylori [86] and insect Drosophila melanogaster [87] have been studied. The common properties exhibited by all of these studied protein interaction networks is- scale free nature, small-world effect and modular organization with motifs (my analysis of the human PPI is in tune with these properties). It was observed that 70 to 80 percent of protein interaction partners share at least one function. It was also found that lethality and degree have a correlation. The importance of proteins can be inferred by analysis of their local topological properties. Jeong et al found a positive correlation between lethality and connectivity [85].

Figure 5.1. Sample Transcription Regulatory Network
The Metabolic networks are the most studied biological networks. Metabolism consists of a set of reactions that are catalyzed by enzymes to produce essential components. These chemical reactions are organized into metabolic pathways, in which one chemical is transformed into another, such a structure can be naturally modeled as a complex network [52]. Two types of elements take part in a metabolism (metabolites and reactions) thus, can be modeled in two different ways - one is where metabolites are the nodes that are linked if they participate in the same chemical reaction; the other is a reaction graph where nodes are reactions and metabolites are the links [38] (see figure A.3). Jeong et al studied the metabolic networks (modeled with nodes as metabolites and the edges as the reaction between them) of 43 different organisms and drew the following conclusions about these networks - the degree distribution of these networks follows a power law with an exponent of 2.2. These networks exhibit small world property with an average path length of 3. They also observed that the diameter for all these networks was
found to be same; but increases logarithmical with the size of the network. The network was found to be modular with hierarchial topology [88].

![Sample Metabolic Network](image)

**Figure 5.3.** Sample Metabolic Network

My main aim in writing this extensive literature review is to -

1. Present an overall review of complex networks.

2. Provide a reasonable insight into networks by discussing the modeling and properties of a wide variety of systems with examples. So that one will be able to model and understand a system as a networks.

### 5.3 Protein Network

Protein-Protein interaction, metabolic, gene expression and signal transduction pathways all fall under the term Cellular Networks. In a Protein-Protein interaction network the nodes are proteins and any interaction between them is represented as an edge. Protein interaction maps have been obtained by the yeast
two-hybrid (Y2H) or the Tandem Affinity Purification followed by mass spectrometry [38]. The Protein to Protein interactions have been reported in a number of databases HPRD, BIND-Bimolecular Object Network Databank [89], DIP-Database of Interacting Proteins, HPRD- Human Protein Interaction Database [90], EBI- European Bioinformatics Institute [91], MINT- Molecular INTeraction database [92] or PIP-Protein Protein Interaction Prediction [45].

These databases are incomplete. The researchers estimate 154,000-369,000 protein interactions in the human network but are only able to establish / prove ten percent of these owing to the high false positive rate [93]. The false positives are usually computed by comparing two databases with each other and to a reference dataset [94]. The completeness, is computed by assay saturation or the dead reckoning method [95]. The network of proteins is a descriptive and predictive model. In spite of the incompleteness these networks have made great contributions. Of all the free online data repositories, the University of Dundee data repository reports the largest number of interactions (37,800) and proteins (69,965) [96]. They also have a unique way of quantifying the protein interaction by a score [97]. In the analysis all the interactions with the reported score greater than or equal to 1 have been considered. The data consists of 8254 unique proteins and 37,754 interactions.
5.4 Network Analysis

The Figure 5.1 is the result of the protein interaction data obtained from the database. To understand this picture or the interaction network we need to perform further analysis. The network topology needs to be analyzed to understand the emergent properties of the network. The emergent and the predictive properties of these networks are very important as the entire network is unavailable for analysis. To identify the topology of the network, the following parameters are computed:

1. Degree distribution
2. Betweeness centrality
3. Clustering coefficient
4. Modularity
5. Assortative mixing

6. Average Path length

The details are discussed in the following sub-sections.

5.4.1 Degree distribution

The spread of the number of edges a node has or a node degree can be characterized by a degree distribution function $P(k)$. This gives the probability that a randomly selected node has a particular number of edges. The degree distributions of all the networks fall into one of the four categories namely-Poisson, Power-law, Exponential or Power-law with an exponential cutoff. The Poisson degree distribution signifies a random graph model; the other three signifies a scale-free graph model.

![Figure 5.5. Degree Distribution.](image-url)
The degree distribution of the protein-protein interaction network (Fig : 5.2) follows a power law ($\alpha = 2.385$) with an exponential cutoff at $k_c = 200$ [85] [98]. This implies that the network is heterogeneous with a few high degree and a few low degree nodes.

5.4.2 Betweenness Centrality

The betweenness centrality of a vertex $i$, is the number of paths between other vertices or nodes of the network that pass through node $i$ [72].

$$g_k = \sum_{i \neq j} \frac{C_k(i,j)}{C(i,j)}$$

(5.6)

Where, $C(i,j)$ is all the shortest paths between node i and node j; $C_k(i,j)$ represents all the paths that pass through node k (Fig:5.3)

Figure 5.6. Network Diagram-Betweenness Centrality.

The slope of the human protein-protein interaction network is -2.71 (Fig:5.4). This implies that the protein-protein interaction network’s betweenness centrality distribution follows the scale free betweenness distribution. It was found that the core protein interaction network of yeast follows a power law betweenness distribu-
tion with a slope of -2.2 [99].

Figure 5.7. Betweenness Distribution.

5.4.3 Clustering Coefficient

The clustering coefficient $C_i$ is the local measure quantifying the likelihood that the neighboring nodes of a particular node $i$ are connected to each other. It is calculated by the ratio between the number of edges $E_i$ that actually exist between these $k_i$ nodes and the total number of cliques $k_i(k_i - 1)/2$ gives the value of the clustering coefficient of node $i$ [73].

$$C_i = \frac{2E_i}{k_i(k_i - 1)} \quad (5.7)$$

For random and scale free networks the clustering coefficient of a node is independent of the degree, i.e. hubs and low degree nodes would have the same clustering coefficient. For the human protein interaction network the clustering
5.4.4 Modularity

Many biological networks naturally decompose into modules. A module can be defined as a group of nodes with large number of edges falling within the group as opposed to leaving the group. There are several algorithms available to discover the modules in a network. We found that there are six modules in the network with a respectable modularity score of 0.543 (Fig 5.6). Nodes of a particular color belong to one module.
A module in the PPI network might imply - all the proteins belonging to a particular module performs a particular task. In other words all the proteins in a particular module work together to accomplish a particular task. This hypothesis needs validation. Guimera and Amaral [102] identified that most of the metabolites in a module belonged to one major pathway. They used simulated annealing for identifying the modules in the Kyoto Encyclopedia of Genes and Genomes (KEGG) dataset and then cross-validated with the KEGG pathway dataset.

![Protein Interaction Network showing the different modules.](image)

**Figure 5.9.** Protein Interaction Network showing the different modules.

### 5.4.5 Assortative mixing

Assortative mixing is a measure that weighs the correlation among degrees in a graph [38]. In other words assortative mixing means the preference of high degree nodes to attach to other high degree nodes and low degree nodes to low degree nodes. Disassortative mixing means that high degree nodes attach to low degree nodes and vice-versa. The Pearson correlation coefficient of the degrees at the two
ends of an edge are used for quantification; value ranges from -1 to 1 [60]. Negative indicates disassortative mixing and positive assortative mixing. The assorative mixing of a network can be computed using the equation 5.3. The assortative mixing of the protein interaction network was found to be -0.07. This implies that the network is dissortative in nature.

\[
r = \frac{M^{-1} \sum_{i,j} j_i k_i - [M^{-1} \sum_i \frac{1}{2}(j_i + k_i)]^2}{M^{-1} \sum_i \frac{1}{2} (j_i^2 + k_i^2) - [M^{-1} \sum_i \frac{1}{2}(j_i + k_i)]^2}
\] (5.8)

5.4.6 Average Path Length

The average of the shortest path length between any two node pairs is called the average path length. The average path length of the protein interaction network is 4.05. The average path length scales logarithmically with the network size (log N/log <k> = 3.27).

5.4.7 Summarizing the Entire Network Properties

The degree distribution and the betweenness centrality follow a power-law distribution. This implies that the network is heterogeneous and contains hubs. The average path length scales logarithmically with the network size (this can be seen from section 5.3.6.). This implies that the network is a small world network. The clustering coefficient scales inversely with the network size. This implies that the network is a hierarchial network [103]. Thus, the protein interaction network is a
<table>
<thead>
<tr>
<th>Graph Attributes</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Degree</td>
<td>13.80151</td>
</tr>
<tr>
<td>Average Path Length</td>
<td>4.05012</td>
</tr>
<tr>
<td>Average Clustering Coefficient</td>
<td>0.3168</td>
</tr>
<tr>
<td>Average Betweenness Centrality</td>
<td>0.00056</td>
</tr>
<tr>
<td>Assortative Mixing</td>
<td>-0.07014</td>
</tr>
<tr>
<td>Modularity</td>
<td>0.543</td>
</tr>
</tbody>
</table>

*Table 5.1. Summarized Network Parameters*

Hierarchial Network (Fig: 5.6).

### 5.5 Protein Ranking

The proteins with top ten percent of the betweenness centrality are considered and then they have been ranked according to the value (highest to lowest). Table: 5.2 shows the list of the top 15 proteins of interest. The betweenness centrality has been used as a measure for ranking, as it is an important measure for keeping the network connected compared to degree [59]. It can be seen from figure 5.3, that node k has a low degree but removal of this node will split the network into two subgraphs. There are 316 Hershey proteins that could be mapped onto the protein interaction network (Dundee dataset). Out of these 316 proteins 96 of them belonged of module 3. That is thirty percent of the Hershey proteins belonged to module 3, we would expect seventeen percent of them to belong to each of the six modules (already mentioned in the 5.3.4. section). As most of the test proteins fall into module 3 this implies that there might be a correlation between the proteins of module 3 and diabetes.
<table>
<thead>
<tr>
<th>UniProt</th>
<th>Frequency</th>
<th>Degree</th>
<th>Betweeness Centrality</th>
<th>Module</th>
</tr>
</thead>
<tbody>
<tr>
<td>P62158</td>
<td>CALM – HUMAN</td>
<td>2</td>
<td>207</td>
<td>0.028753098</td>
</tr>
<tr>
<td>P61769</td>
<td>B2MG – HUMAN</td>
<td>2</td>
<td>110</td>
<td>0.020308282</td>
</tr>
<tr>
<td>P00734</td>
<td>THRBB – HUMAN</td>
<td>6</td>
<td>190</td>
<td>0.016823405</td>
</tr>
<tr>
<td>P62826</td>
<td>RAN – HUMAN</td>
<td>2</td>
<td>115</td>
<td>0.008302618</td>
</tr>
<tr>
<td>P00747</td>
<td>PLMN – HUMAN</td>
<td>6</td>
<td>107</td>
<td>0.007041897</td>
</tr>
<tr>
<td>P62805</td>
<td>H4 – HUMAN</td>
<td>2</td>
<td>126</td>
<td>0.006851843</td>
</tr>
<tr>
<td>P00742</td>
<td>FA10 – HUMAN</td>
<td>2</td>
<td>137</td>
<td>0.006550363</td>
</tr>
<tr>
<td>P26038</td>
<td>MOES – HUMAN</td>
<td>2</td>
<td>57</td>
<td>0.004757423</td>
</tr>
<tr>
<td>P07358</td>
<td>CO8B – HUMAN</td>
<td>4</td>
<td>56</td>
<td>0.004365211</td>
</tr>
<tr>
<td>P55209</td>
<td>NP1L1 – HUMAN</td>
<td>2</td>
<td>66</td>
<td>0.004217259</td>
</tr>
<tr>
<td>P02671</td>
<td>FIBA – HUMAN</td>
<td>6</td>
<td>58</td>
<td>0.003666837</td>
</tr>
<tr>
<td>P07237</td>
<td>PDIA1 – HUMAN</td>
<td>2</td>
<td>16</td>
<td>0.003423083</td>
</tr>
<tr>
<td>P55083</td>
<td>MFAP4 – HUMAN</td>
<td>2</td>
<td>4</td>
<td>0.003304903</td>
</tr>
<tr>
<td>P01005</td>
<td>ANT3 – HUMAN</td>
<td>4</td>
<td>96</td>
<td>0.002868537</td>
</tr>
<tr>
<td>P55072</td>
<td>TERA – HUMAN</td>
<td>2</td>
<td>32</td>
<td>0.002805883</td>
</tr>
</tbody>
</table>

Table 5.2. Top 15 Proteins of interest
Chapter 6

Conclusions and Future Work

6.1 Introduction

This chapter is focused towards discussing conclusions from both the methods of implementation and the future work. The first method discussed in this thesis was the data mining approach (refer to chapter 4) and the second the network mining approach (refer to chapter 5).

6.2 Conclusion

In this thesis two methodologies were used to identify the significant proteins that cause diabetes. The first method is data mining approach which has been discussed in Chapter 4, and the second one based on the network science approach has been discussed in Chapter 5. The datamining approach used was similar to the popular Gene Filtering method which consisted of screening and ranking. The test
proteins (proteins obtained from the Hershey Medical Center) obtained from the Protein Pilot were first classified into three classes namely- Class 1 (Up-regulated), Class 2 (Down-regulated) and Class 3 (no change) depending on the protein ratio value. The proteins of interest were that of Class 1 and Class 2 refer to Chapter 4 for further details. These proteins were ranked based on their occurrence on the Pareto-Front and the frequency (number of patients in which the protein is present). The obtained top eleven proteins of interest are shown in the table 4.1.

The network science approach is a more holistic approach. In this the Protein-Protein Interaction (PPI) network was generated using the data from the Dundee dataset. After performing alot of analysis, the network was found to be Hierarchial in nature (methodology used and details are discussed in Chapter 5). The network parameters like degree, betweenness centrality, average path length, assortativeness and clustering coefficient were obtained. The location and other parameters of the test proteins were obtained. Then, the top ten percent of the betweeness centrality (reason discussed in chapter 5) proteins were ranked according to the value. These are the top fifteen proteins of interest.

In the network science approach only 316 of the 595 test proteins could be mapped onto the network. The reasons being-

1. Few of the test proteins were lost due to the conversion; the obtained test proteins were in the UniProt format and the proteins from the Dundee dataset were in the International Protein Index (IPI) format. The test proteins were
Table 6.1. Top 15 Proteins of interest

<table>
<thead>
<tr>
<th>UniProt</th>
<th>Frequency</th>
<th>Degree</th>
<th>Betweenness Centrality</th>
<th>Module</th>
</tr>
</thead>
<tbody>
<tr>
<td>P62158</td>
<td>CALM – HUMAN</td>
<td>2</td>
<td>207</td>
<td>0.028753098</td>
</tr>
<tr>
<td>P61769</td>
<td>B2MG – HUMAN</td>
<td>2</td>
<td>110</td>
<td>0.020308282</td>
</tr>
<tr>
<td>P00734</td>
<td>THRBI – HUMAN</td>
<td>6</td>
<td>190</td>
<td>0.016823405</td>
</tr>
<tr>
<td>P62826</td>
<td>RAN – HUMAN</td>
<td>2</td>
<td>115</td>
<td>0.008302618</td>
</tr>
<tr>
<td>P00747</td>
<td>PLMN – HUMAN</td>
<td>6</td>
<td>107</td>
<td>0.007041897</td>
</tr>
<tr>
<td>P62805</td>
<td>H4 – HUMAN</td>
<td>2</td>
<td>126</td>
<td>0.006851843</td>
</tr>
<tr>
<td>P00742</td>
<td>FA10 – HUMAN</td>
<td>2</td>
<td>137</td>
<td>0.006550363</td>
</tr>
<tr>
<td>P26038</td>
<td>MOES – HUMAN</td>
<td>2</td>
<td>57</td>
<td>0.004757423</td>
</tr>
<tr>
<td>P07358</td>
<td>CO8B – HUMAN</td>
<td>4</td>
<td>56</td>
<td>0.004365211</td>
</tr>
<tr>
<td>P55209</td>
<td>NP1L1 – HUMAN</td>
<td>2</td>
<td>66</td>
<td>0.004217259</td>
</tr>
<tr>
<td>P02671</td>
<td>FIBA – HUMAN</td>
<td>6</td>
<td>58</td>
<td>0.003666837</td>
</tr>
<tr>
<td>P07237</td>
<td>PDIA1 – HUMAN</td>
<td>2</td>
<td>16</td>
<td>0.003423083</td>
</tr>
<tr>
<td>P55083</td>
<td>MFAP4 – HUMAN</td>
<td>2</td>
<td>4</td>
<td>0.003304903</td>
</tr>
<tr>
<td>P01008</td>
<td>ANT3 – HUMAN</td>
<td>4</td>
<td>96</td>
<td>0.002868537</td>
</tr>
<tr>
<td>P55072</td>
<td>TERA – HUMAN</td>
<td>2</td>
<td>32</td>
<td>0.002805883</td>
</tr>
</tbody>
</table>

converted into the IPI format (details are discussed in Chapter 5).

2. All the complete human interactome has still not been deciphered. This is the reason why most of the test proteins were not mapped onto the PPI.

3. In network science, analysis is performed on the largest connected component. Much or nothing can be said about the proteins that are not present in the largest connected component.

Inspite of few proteins not being mapped onto the network the network science approach is a much better approach as it gives us a holistic picture of the entire test proteins - their location, connectivity and other properties of interest.
6.3 Future Work

In this thesis we tried to identify the significant proteins that cause diabetes. A disease or a function is rarely performed by one molecule it is usually the result of interaction between a number of different molecules. Therefore, one cannot just look at proteins or genes or metabolites and identify the drug targets to state that these are important; one needs to look at the genes, proteins and metabolites as well as their interactions to identify the drug targets.

Chapter 4 and 5 describes our approach and the significant proteins that were identified. This work has a lot of potential and the following are a few ideas for future work-

• **Complete Understanding of the Protein Interaction Network:**

  The analysis in this thesis only provides a key hole view into the protein interaction network. Further analysis needs to be performed to understand the entire network and its dynamics by using tools from complex networks.

• **Adding the protein ratio value to the current analysis:**

  In the holistic complex network based systems biology approach we have only used the protein names; protein ratio values have not been used. Further analysis can be performed in the same lines by identifying the - links between the protein ratios and the modules, protein ratios and the node property and other analysis on these lines. This analysis may reveal new relationships and
new significant proteins.

• **Drug target Identification:**

Future work can be focused towards obtaining the drug targets. To get at the drug targets we might need to look at a few more parameters and validate the current results. To validate the results statistically as well as biologically we need to increase the size of the dataset.

• **Signalling Pathway Analysis:**

Better understanding of diabetes can be obtained by understanding the signalling pathways of the disease. To analyze or understand the signalling pathways one needs to look at the complete omic data and also the components at the cellular level.

• **Correlation of diabetes with other diseases:**

Future work can be focused towards uncovering the relationship between obesity, diabetes and the other comorbid conditions. Research till date has been focused towards identifying the correlation between the different diseases; current researchers are working towards understanding why there is a correlation.


[91] **DATABASES, E.-E.-P.** URL http://www.ebi.ac.uk/IPI/IPIhelp.html


