A NOVEL NETWORK BIOLOGY APPROACH TO DRUG TARGET SELECTIONS

Ragini Pandey

Submitted to the faculty of the School of Informatics

in partial fulfillment of the requirements

for the degree of

Master of Science in Bioinformatics,

Indiana University

December 2008

Accepted by the Faculty of Indiana University,

in partial fulfillment of the requirements for the degree of Master of Science

in Bioinformatics

Master's Thesis Committee

_____

Dr. Jake Yue Chen, Ph.D., Chair

_____

Dr. Narayanan Perumal, Ph.D.

_____

Dr. Pedro Romero, Ph.D.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor, Dr. Jake Yue Chen, for having faith in me and giving me the opportunity to join his research team. I am grateful to him for introducing me to scientific research and all that it entails. I also appreciate and extend my thanks to him for playing a major role in honing my writing and presentation skills and for the constant support and guidance he gave me as I worked to accomplish this project.

I also thank my committee, Dr. Narayanan Perumal and Dr. Pedro Romero, for their valuable feedback and support. I wish to thank all the members of my lab for being cooperative, encouraging and supportive throughout the project. I especially thank Jiao Li for her help in presenting my work and reviewing my thesis and Dr. Scott Harrison for the support and help he gave me throughout the process and being available whenever I needed. I would also like to thank Sudhir R. Chowbina for taking some time off his work to review my thesis. A special thanks to all other members who helped either directly or indirectly with my research.

I extend my warmest thanks to the School of Informatics for equipping me with the tools and technologies required for the successful completion of my research and for the wonderful academic life at IUPUI. My deepest is gratitude to my husband, my parents and my loving family; I could not have become what I am today nor achieved what I have achieved without them. Last, but not least, I thank all my friends and relatives who have believed in me and have been a source of great moral strength.

ABSTRACT

Conventional drug discovery focuses on single protein targets and follows a "sequence, structure, and function" paradigm for selecting best protein targets to screen lead chemical compounds. This established paradigm simply avoids addressing directly the challenge of evaluating chemical toxicity and side effects until a later stage of drug discovery, resulting in inefficiencies and increased time and cost. We developed a new "network biology" perspective to assess proteins as potential drug targets using emerging biomolecular network data sets. To do so, we integrated several types of biological data for current drug targets from DrugBank, protein interaction data from the HAPPI and HPRD databases, literature co-citation data from PubMed, and side effects data from FDA-approved drug usage warnings. We used the Bayes factor and Positive Predictive Values to examine the use of certain network properties, such as network node degrees and essentiality, to predict candidate drug targets. We also developed a metric to evaluate a protein target's overall side effects by taking into account aggregated side effect scores of all FDA-approved drugs targeting the protein. We discovered that non-essential protein with lower-to-medium network node degree could better serve as drug targets when combined with conventional protein function information. Integrated biomolecular associations, instead of physical interactions, are better sources for predicting drug targets with network biology methods. Our network biology framework presents exciting promises in developing better drug targets that lower the side-effects at later stages of drug development and help establish the field of "network pharmacology."

TABLE OF CONTENTS

PAGE

LIST OF FIGURES

PAGE

CHAPTER ONE

INTRODUCTION: BASIC CONCEPTS AND THESIS OVERVIEW

1.1 Basic Concepts: Network Biology and Drug Discovery

Use of the Network Biology perspective in drug discovery would help in determining the way potential targets are most likely to behave, their interconnections, and the possible knock-out effect of interfering with them. Network Biology is an assembly of interconnected functional modules that integrate and coordinate the cell's major biochemical activities and responses to external and intrinsic signals. The theory predicts that modulating multiple nodes simultaneously is often required for modifying phenotypes[1]. The concept includes network topological properties, such as network centrality, clustering coefficient; network functional properties, such as lethality/essentiality; and dynamical properties, such as entropy, fractal, robustness and complexity.

The most critical step in the drug discovery process is target identification and validation. Current paradigm of drug discovery is single target[1-3] or selective target[4]. A drug target is a key molecule involved in a particular metabolic or signaling pathway that is specific to a disease condition or pathology, which could be modified by an external stimulus. The toxicity, specificity and the inability to obtain potent compounds against polysaccharides, lipids and nucleic acids limit the domain of drug targets mostly to proteins[5]. Some of the properties which differentiate a protein as a candidate drug target are the ability to bind to small molecules (structural property, Figure 1.1), overlap with disease[5], connectivity (the number of other proteins with which it interacts) and

between-ness (shortest path between two networks) (Figure 1.1). These properties make it important to study them from the network biology perspective. Though being a single targeted drug discovery, the drugs have been reported to interact with several other proteins.

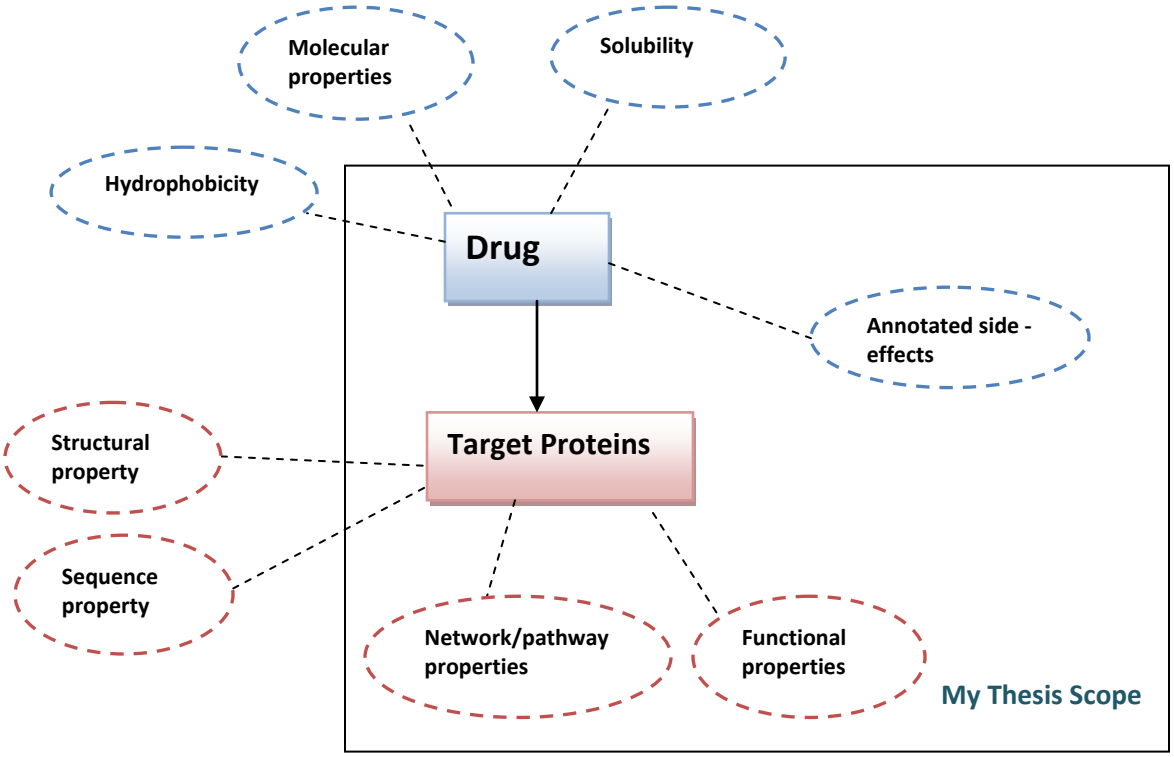


Figure 1.1: Overview of relationship between main entities of the thesis (rectangle showing the scope of the thesis)

For a drug to be effective it must bind to its target protein with a reasonable degree of potency[5]. The other properties include compliance with the 'Rule of five' (necessary molecular properties), solubility, and hydrophobicity (Figure 1.1). Drugs often cause side effects which have been accredited to a number of molecular scenarios (Figure 1.1) including the interaction with the primary or additional targets (off-targets binding), downstream pathway perturbations, kinetic and dosage effects, and drug-drug

interference. Keeping these scenarios in mind while selecting targets would help minimize the side effects. But the current paradigm of single targeted drug discovery does not seem to consider the different properties of drug target proteins. To target proteins which are a part of complex networks, network study is of utmost importance.

Protein interactions can be modeled as a network of nodes or components. The edges in the network are the physical and the functional interactions among the proteins. The number of interacting partners a protein has determines its Degree of Connectivity, which in turn classifies a protein to be a hub or a non-hub with the hubs being proteins that are highly connected. The node connectivity (degree) follows a power law distribution; i.e., there are a small number of nodes with a large number of connections and a large number of nodes with a small number of connections[6]. These hubs are quite stringent to random deletions. Based on the partners and the time and location of interaction, the hubs are further classified as date hubs and party hubs[7].

Essentiality is one of the properties (Figure 1.1) which refers to a protein that when knocked out renders the cell unviable. It is the behavior of a network on deleting a protein which defines it to be essential or non-essential (Centrality and lethality rule [6]). However, there are controversies among the structural importance[6] and the functional importance[7] of a protein in a network which decides its essentiality. But to date, there is no concrete definition as to what constitutes essentiality.

The aim of this study is to do statistical analysis on existing drug targets for their network properties, such as network node degrees and essentiality, with the intention of finding new characteristics of drug target proteins from the network perspective that would help in predicting candidate drug targets. Apart from high level analysis of drug

targets, this study analyzes in depth the network properties of drug target proteins, which makes them stand out in the crowd of potential targets. Finding specific node degrees of drug target proteins will help cut down on the large number of proteins out there and also would introduce new potential targets that have the characteristics for becoming good targets causing minimal side effects. The analysis of whether essentiality should play a role, or whether it has any role, in target selection or in causing side effects by drugs could bring a new prospective of target selection. The side effect analysis will help in determining which could be future good targets with minimal side effects from the existing targets.

1.2 *Motivation*

Current paradigm of drug discovery focuses on finding a 'magic bullet'[4] for a disease with minimal side effects. But it avoids addressing directly the challenge of evaluating chemical toxicity and side effects until a later stage of drug discovery, resulting in inefficiencies, as well as increased time and cost, which ultimately results in fewer drug approvals. It also does not seem to work for complex disorders that progress by multiple pathways and mechanisms. Another problem is the abundance of 'me too' drugs or the same drug target protein being targeted over and over by different drugs for different diseases[2].

Owing to the complexity of organisms and the complex interactions its cells have to undergo in performing several activities, the study of potential targets in relation to intercellular network of interactions becomes a requirement. Today, the biggest challenge in drug discovery is the high attrition rate. Many drugs fail due to the poor understanding of the system's molecular functions they have to target and the way they might affect the

entire system's network[8]. Using the "network biology" perspective to proteins as potential drug targets would be a solution to the problem.

1.3 Thesis Contribution

1. We developed a network-topology-based method that would facilitate the evaluation of potential drug targets.

   We developed a novel "network biology" approach to select potential drug targets. This method shows a new way to select targets by using the network properties such as network node degrees of a protein.

2. We developed a method for evaluating the true goodness of the drug targets.

   Through our work we show that simply selecting a target does not qualify it to be a 'good target' that would cause minimum side effects. We show how to select a 'good target' by combining the network property of protein to the side effects caused by drugs.

3. Our method of rating drugs based on side effects would assist in standardizing drugs.

   We developed a method that rates all drugs depending on the side effects they cause.

4. We developed a metric to evaluate drug targets based on the side effects which will provide an idea of the severity of the side effects they would cause when targeted.

   We show a way to evaluate the target protein's overall side effects by taking into account the aggregated side effect scores of all FDA-approved drugs targeting the protein.

1.4 *Organization of the Thesis*

This thesis is divided into five chapters. The first chapter deals with the introduction and overview and Chapter 2 discusses the literature reviews in the context of hubs, protein essentiality and drug targets. Chapter 3 describes the methods and formulae used in the study for the statistical network analysis of drug targets, as well as the data collection and the challenges faced in doing so. Chapter 4 analyzes the graphs and presents the results and Chapter 5 concludes the thesis with a discussion.

CHAPTER 2

LITERATURE SURVEY

This chapter introduces topics that are necessary for understanding the problem and analyzing its results. Starting with the role of systems biology in drug discovery, the chapter proceeds to explore the current approaches for the process while highlighting the use of network biology in selecting candidate drug targets. The chapter ends with the problem definition and related research questions.

2.1 Drug Discovery

Drug discovery is the process by which drugs are discovered and/or designed. The most critical step in the process is the **target** selection**,** a key molecule involved in a particular pathway specific to a disease condition and possessing the potential of being modified by an external stimulus. Potential targets may, for example, be proteins whose genes are over-expressed or those associated with the defective proteins but are themselves unsuitable to be targeted by a small molecule. Drugs have to bind to these potential target proteins to cure a disease condition. But though crucial to the process, target identification is largely based on circumstantial evidence where a protein is selected based on its references to a disease. Traditionally, to study biology and human health, individual proteins and genes were investigated one at a time in order to understand their functionality and contribution toward a specific functional aspect of the organism. But this ignores the complexity of the multi-cellular organisms, which leads to a limited understanding of the human body and its operation and, thus, limits the capability to best predict, prevent, or remedy potential health problems.

2.2 *Systems Biology in Drug Discovery*

"Systems biology allows you to understand how a very dynamic system works so that you can find ways to target disease better than just blocking a pathway," said Howard Schulman, vice president of research at California-based SurroMed.

Systems biology forms a cycle of experiments performed, result analysis and novel knowledge (Figure 2.1).

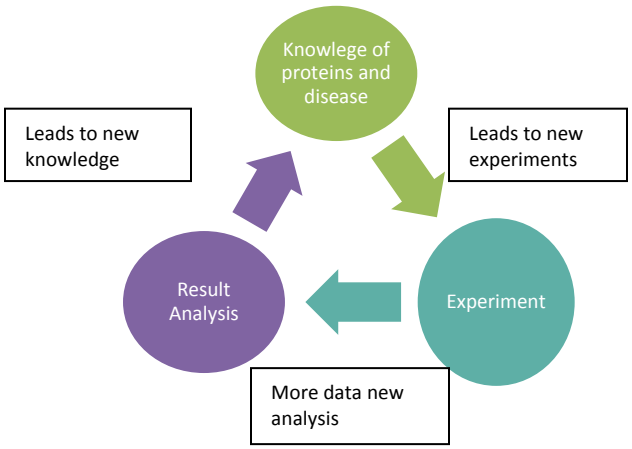


Figure 2.1

- In order to understand the complexity of the biological systems, it is necessary to perform experiments which in turn generates huge amount of data.
- The data generated from experiments would be of no value without analysis. Thus computation analysis becomes crucial for the vast data generated by experiments. As a result, we gain new knowledge.
- The new knowledge obtained as a result of analysis of new data is then used to perform new experiments, and, as such, the cycle continues.

Drug discovery and systems biology began as traditional folk medicine[9, 10] where compounds were studied, experiments were performed and the results were used to

perform new experiments. Today, systems biology deals with understanding physiology and disease at the molecular pathway level; within regulatory networks, cells, tissues, organs; and ultimately the whole organism. An Omic approach to systems biology focuses on genes, proteins, and metabolites. The drug industry also has started to implement these Omic approaches to complement traditional approaches of target identification in generating hypotheses and for experimental analysis. Advances in systems biology suggest that complex diseases may not be effectively treated by interventions at single nodes only[1]. Two central problems of drug discovery today are:[11]-

- the search for disease-related targets
- the study of drug–protein interactions and protein–protein interactions

With the huge amount of data produced from genome sequencing, systems biology promises to help solve the problems by uncovering new drug targets.

Hopkins et al introduced the concept of 'Druggability' in the paper, 'The druggable genome'[5]. The authors revealed that out of four types of macromolecule, namely lipids, nucleic acids, polysaccharides, and proteins (which can be interfered with using small-molecule therapeutic agents), it is mainly **proteins** that **can be used as targets.** But neither all proteins are 'druggable,' nor can all be targeted by drugs; only proteins that have the ability to bind and also overlap to a disease can be targeted.

2.3. Current Approach to Drug Discovery

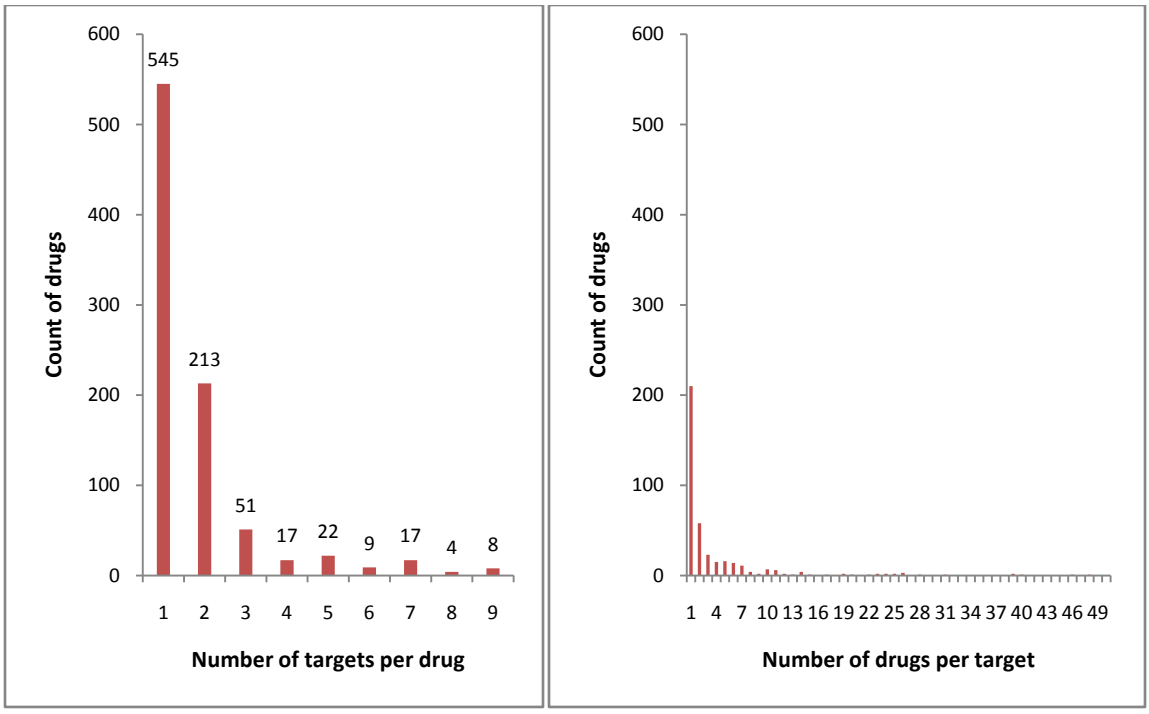
Drug discovery today mainly focuses on the single-target approach: 'One target, one drug and one disease'[1] with the drug acting as a selective 'key' that fits into the 'lock' of

a specific drug target. Despite the fact that the target-based paradigm eases certain development activities, it also has certain drawbacks, as follows:

- It isn't always guaranteed to work. It might not always affect complex systems in the desired way even if it completely changes the behavior of its immediate target.[8]
- Certain drugs work only for a certain population of patients. – Iressa *works only within a particular population. It gives a good response in only 1/10 of the patients who receive it [3, 12]*.
- It ignores the enormous complexity of cells and tissues. For example, single targets might have 'back-up' systems which could sometimes be different and may not respond to the same drug[8].
- It limits the ability of researchers to identify innovative targets and/or mechanisms of action (including combination therapies) by limiting druggable space to recognized targets and modalities[9].
- It usually cannot fight multi-genic diseases such as cancer, or diseases that affect multiple tissues or cell types. Cancer and other nervous system defects are deregulation of many biochemical pathways.

However, the fact that the drugs bind to several other targets is often ignored. For example, *Gleevac* which is developed using a single target approach actually works by attaching to a key part of an overactive protein that causes chronic myeloid leukemia *[4]*. While studying antipsychotic drugs, it was found that the successful drugs actually act on multiple, rather than single, targets.

Yildirim, M.A., et al[2], in their 'Drug-Target Network' study, developed a network of current drugs and their targets. Through their analysis of the network they revealed that though maximum drugs have only one target, there are several drugs which have multiple targets (Figure 2.2(a)) and many drug target proteins are, in turn, being targeted by more than one drug (Polypharmacology) (Figure 2.2(b)).



(a) (b)

Figure 2.2 (a) Number of targets per drug (b) Number of drugs per target

The analysis of a drug-target network also showed the abundance of `me too ` drugs, meaning the old targets were being targeted again and again[2]. When mapping the drug-target network to a human disease-gene network, it was found that not only did the drugs act on multiple targets, but the targets also were involved in multiple diseases[1]. Thus, although the current approach of drug discovery may be single targeted, behind the scenes the drugs also act on other targets and those targets are not

being considered. It has also been found that 30% to 40% of drugs fail in clinical trials because of inappropriate target selection[5]. This clearly indicates that the current paradigm of drug discovery needs to be revisited.

Another market study on the rate of drug approvals showed that although there is a continuous increase in the cost per experiment performed, the rate of drug approvals are on declining, which should not be the case, given the enormous increase in resources and advancement in the research technologies.

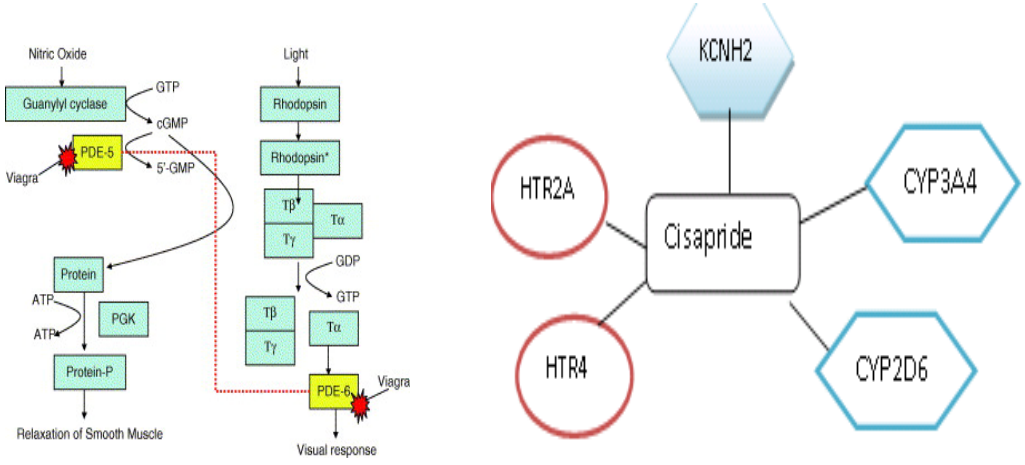
Viewing drug action through the lens of network biology may provide insights into improving drug discovery for complex diseases[1]. The study of network biology in relation to targets becomes of utmost importance considering the complexity of cells and tissues and the fact that drugs act on multiple targets at the same time**.** Hopkin's et al. in their 'Network Pharmacology' study, mapped the drug target network on the protein interaction network to reveal all the proteins be targeted by the same drug and their interconnection. They stressed the fact that drug efficacy and toxicity can be well understood by action at specific nodes and hubs.

2.4 Drugs and their Side Effects

Side effects are harmful and undesired results from a medication or other intervention. They are complex phenomenological observations that have been attributed to a number of molecular scenarios, including interaction with the primary or additional targets (off-targets hereafter), downstream pathway perturbations, kinetic and dosage effects, and drug-drug interference. After a drug enters a cell, it can either interact directly with a receptor or evoke beneficial or detrimental responses by either up- or down-regulation of receptor-activated signaling pathways. Or the drug can undergo

metabolism to products that, in turn, react with the receptor. The majority of these products are inactive (detoxification), but some are reactive and bind[13].

Of all causes, off- target binding is the most important. It has been found that Viagra was designed to target PDE-5 and promote the relaxation of smooth muscle, but the compound also binds to the homologous PDE-6 in the eye, which leads to a "blue vision" side effect in patients[14] (Figure 2.3a). This shows the necessity of studying the targets relevant to the network. **The network approach examines the effect of drugs in the context of a network of relevant protein-protein interactions.**[8] This could help minimize severe side effects. While doing the network analysis of serotonin receptor (HTR4 and HTR2A) with its agonist cisapride, Kuhn et al[15] found that cisapride also binds to the cardiac ion channel hERG (KNCH2) (Figure 2.3b), which leads to arrhythmias as a side effect. In addition, the network also showed the interaction of cisapride with Cytochrome P450 enzymes (CYP3A4 and CYP2D6).



(a) **(b)**

Figure 2.3: (a) Example of off-target binding by Apic, G. et al. (b) Sources of interaction shown by color experiments (red) and text mining (blue)

However, due to limitations related to the availability of data, not much has been done toward studying the co-relation of drugs, targets and their side-effects on a large scale. **It is most likely that novel associations between drugs and protein targets could be found by comparing the side effects profiles of drugs**[15].

2.5 Protein Networks

Proteins are often referred to as the molecular workhorses of the cell since they are responsible for the majority of functions within a living cell. Protein networks help define individual proteins within the context of all other cellular proteins[16]. They form a network of interacting proteins where components or nodes (proteins) are connected by physical and functional interactions (edges)[6]. Based on the degree of connectivity, the proteins are divided into hubs and non-hubs with hubs being the proteins having a large number of connecting partners. However, there is no clear cut-off degree for differentiating a hub from a non-hub. Some researchers have defined them as proteins with more than 5 interacting partners (degree >= 5) [6, 17], whereas others have defined them as proteins that are in the top 20% of the degree distribution (i.e. that have the 20% highest number of neighbors)[18]. The node connectivity (degree) follows a power law distribution, i.e. there are a small number of nodes with a large number of connections and a large number of nodes with a small number of connections[6]. According to the 'Centrality and Lethality Rule' (see introduction), there exists a co-relation between a node's structural importance in the PPI network and its functional importance[6]. The network analysis has shown that the removal of a hub increases the proportion of unreachable pairs in the network (network diameter)[19], and thus increases the mean shortest path between all pairs of reachable nodes in the network.

Haiyuan et al[18] defined network nodes that have many "shortest paths" going through them as 'Bottlenecks'(proteins with a high betweenness centrality). These were believed to control most of the information flow in the network, representing the critical points of the network. *Han et al*[7] further divided the hubs into 'date hubs' and 'party hubs' based on the location and their interacting partners. Date hubs bind to different partners at different times or locations. They organize the proteome, connecting biological processes, or modules, whereas party hubs interact with most of their partners simultaneously. They function inside modules. *Haiyuan et al* believed that bottlenecks with high degrees should have a higher tendency to be date hubs[18]. While studying the characteristics of hubs and non-hubs *Ekmann et al* [20] found that hubs actually are multi domain long proteins which differentiate them from non-hubs, whereas long disordered regions in date hubs help them in flexible binding.

Studying the position of proteins in the biological networks also could be helpful in drug target selection[1]. When mapping the drug-target network to the human protein interaction network, it was found that drug targets tend to have more interactions than average proteins but fewer as compared to essential proteins [1, 2]. The mapping of drug targets to the interaction network also would reveal all the proteins targeted by the same drug and their interconnection. This would be helpful in understanding how the drug reacts at specific nodes and would in turn be helpful in understanding the efficiency and the toxicity of the drug[1]

2.6 Essentiality/Lethality

Essential/Lethal proteins are the proteins which, when knocked out, render the cell unviable. Despite numerous studies performed, it is still unclear as to what causes a

protein to be essential or non-essential. According to *Jeong, H., et al[6],* hubs tend to be essential. In their study they indicated that there exists a correlation between a node's structural importance in the PPI network and its functional importance. According to them, highly connected proteins with a central role in the network's architecture are three times more likely to be essential than proteins with only a small number of links to other proteins.

However *He X, Zhang[17]* challenged their results by saying that the essentiality of a PPI does not seem to be determined by network structures but rather by the particular functions of the interaction. According to them, proteins linked by an essential interaction must be essential, whereas an interaction between essential proteins (IBEP) may or may not be essential. They further demonstrated that betweenness and closeness measure the centrality of a node in the global network structure. *Haiyuan et al[18],* in their study, revealed that 'bottlenecks'(proteins with a high betweenness centrality) tend to be essential. On finding a correlation between degree of connectivity and betweenness they went further to analyze a better predictor of essentiality among two. It was found that betweenness was a better predictor for regulation networks, whereas degree was a better predictor in interaction networks. *Tew et al* [21], in their study, revealed that although the essentiality/lethality of a protein could be found based on the topological position in the network, lethality correlates more strongly with its "functional centrality" than pure physical interaction-based centrality. Goh et al[22], while studying a network of human genes and diseases, found that, although essential, human genes are likely to encode hub proteins, but that the majority of disease genes still are nonessential and show no tendency to encode hub proteins. Their expression pattern indicates that they are localized

in the functional periphery of the network. Apic et al[14] indicated that essential hub proteins have the potential to become future drug targets, for instance, in cancer research.

2.7 Problem Statement and Research Questions

Existing studies have indicated that there is a need to change the current strategy of target selection in drug discovery that is not successful in combating all types of diseases. A continuous decline in drug approvals indicates an immediate need to update the characteristics of candidate drug targets. The importance of the network biology prospective in drug discovery has been recognized by some researchers but is still not fully implied. Hubs are being indicated as being better targets but the definition of a hub is still undecided or unclear. Just indicating hubs as targets opens up a large domain of network node degrees for candidate drug targets. There is a need to go into specific details on what makes the drug target proteins stand out in the crowd. Essentiality is another important network characteristic of a protein but whether it is really considered during target selection or what affect it has on the side effects of drugs is still unknown.

In my effort to address some of the problems above, I integrated the drug target, drug side effect and protein interaction networks to study the underlying relationships (Figure 2.4).

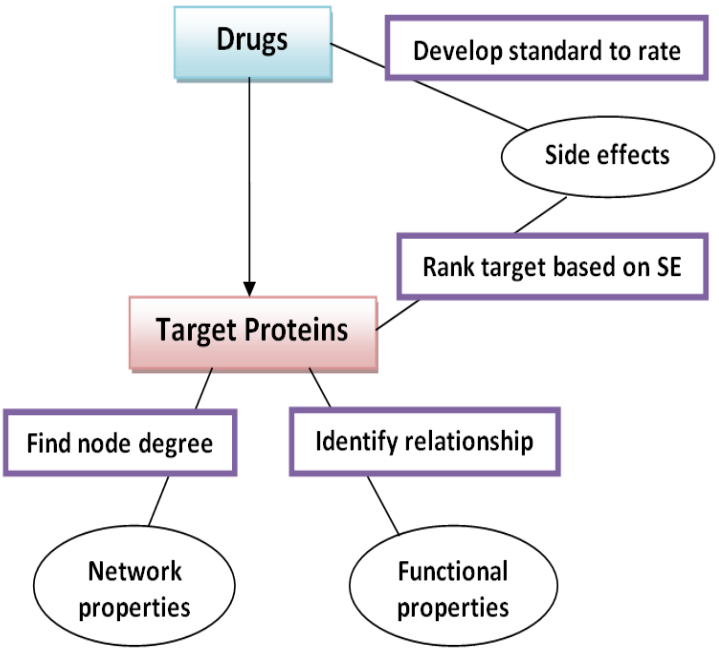


Figure 2.4: Focus points of the thesis

Some questions that will be addressed in the study (Figure 2.5) are:

- Are network hubs/non-hubs preferentially used at drug targets?
- Which network connectivity range of proteins could be better candidates for targets?
- Does the selection of a protein as a target qualify it to be ‘good target’?
- Will network analysis help in finding ‘good targets’?
- Does essentiality play a role in drug target selection?
- Is there a relationship between the essentiality of targets and the side effects of drugs?
- Is it possible to find ‘bad targets’ by analyzing drugs and their side effects?

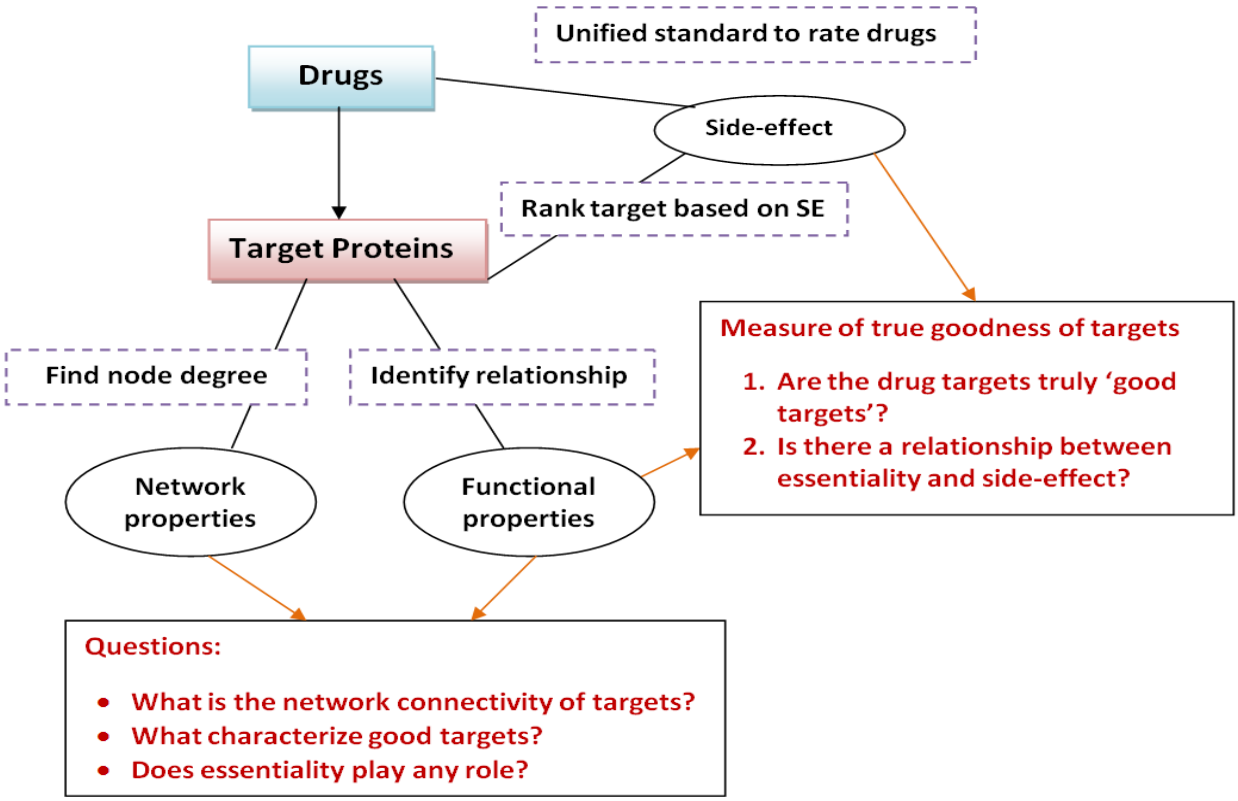


Figure 2.5: Thesis questions

In this study we use the network biology approach to analyze current drugs, their targets and drugs side effects, with emphasis on network properties, such as network node degrees and essentiality. The idea is to study the trend in successful drugs, intending to discover network characteristics that might be helpful in the drug target research and may have been overlooked causing the decline in drug approvals. This would also bring forward new proteins of interest which may not have been considered before but could be good candidates for drug targets. Intermingling the side effects and the network properties would help us find characteristics of potential drug targets with greater probability of causing minimal side effects if targeted.

CHAPTER 3

MATERIALS AND METHOD

The previous chapter discussed the literature related to drugs and drug target proteins, which leads to questions most likely answerable through network analysis. The chapter gave an overview of studies performed in relation to drug targets, protein-protein interactions and the essentiality of proteins. This chapter describes in detail the data sources of the bio-molecular network data sets used in the study, along with the method followed to analyze those data sets with a "network biology" perspective to assess proteins as potential drug targets. The objective of the approach is to find promising characteristics in developing better drug targets that minimize side effects. This would also bring forth a new perspective of analyzing drugs in connection with their targets and the side effects.

Figure 3.1 shows the data integration framework of the study. It gives the overview of all the data sources and the type of data acquired from them and how they were linked to perform the study.

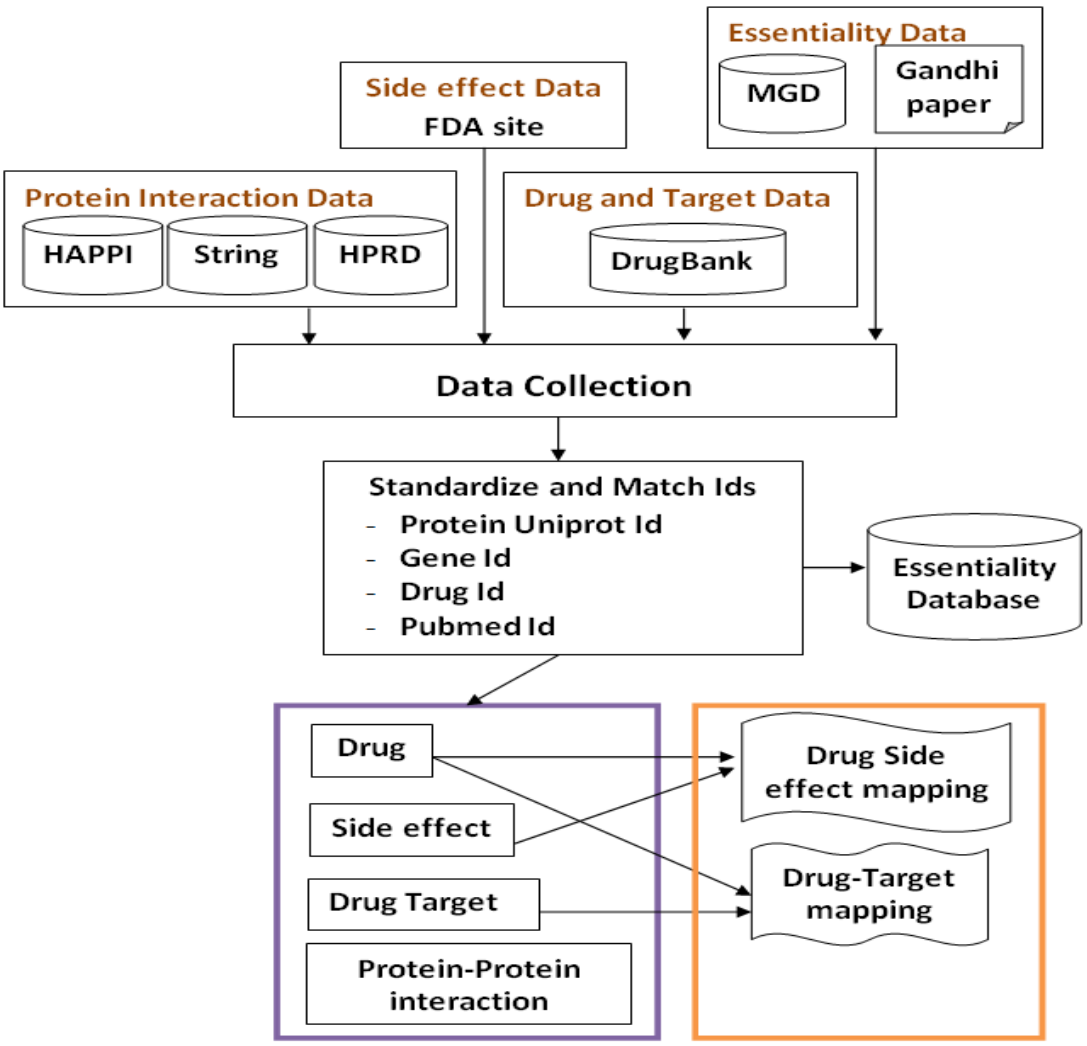


Figure 3.1: Data Integration Framework

3.1 Drug and Drug Target Data

3.1.1 Drug and Target Data Collection

Drug data was downloaded from DrugBank (*http://www.drugbank.ca/*) in 2006[23]. DrugBank is a web-enabled, searchable and comprehensive source that combines detailed drug data (i.e. chemical) with comprehensive drug target (i.e. protein) information. There are more than 4,100 drug entries that include FDA-approved small

molecule drugs, biotechnology drugs and experimental drugs. More than 30 fields describe each drug, which details such information as its generic name, indication, drug type, molecular weight, toxicity, etc. Drug targets are described with their target name, gene name, Uniprot ID, etc. The drug target proteins map to the drugs by the DrugBank accession number. Initial analysis of the data revealed that there were a few Drugbank accession numbers that did map to drug target proteins and there were a few of them where the drug fields were not properly annotated (see appendix). These particular cases were discarded. Uniprot ID was decided to be the universal ID for representing the drug target proteins in the study, so only those drug targets that were identified by the Uniprot IDs were considered in this study. We created a mapping table in order to map the drugs to their targets (identified by Uniprot IDs) (Figure3.1).

3.1.2 DrugBank Data Details

Of the 2,394 drug target proteins identified by Uniprot ID, 2,377 could be mapped to drugs. The 2,394 targets consist of 842 human and 1,552 non-human proteins. Of the 4,247 total drugs, only 3,859 could be mapped to targets; those consisted of 1,062 FDA-approved drugs and 2,797 Experimental Drugs (Figure 3.2). An initial analysis of the data showed most of the drugs had only one target, although there also were a significant number of drugs targeting multiple proteins (Figure 2.2(a)). In the same lines, it was found that although most of the proteins were being targeted by only one drug, there were some which were targets of more than one drug (Figure 2.2(b)). DrugBank also annotates the target proteins with the essentiality information. Of 532 unique proteins with essentiality info in DrugBank, 32 were essential human proteins (Figure 3.6).

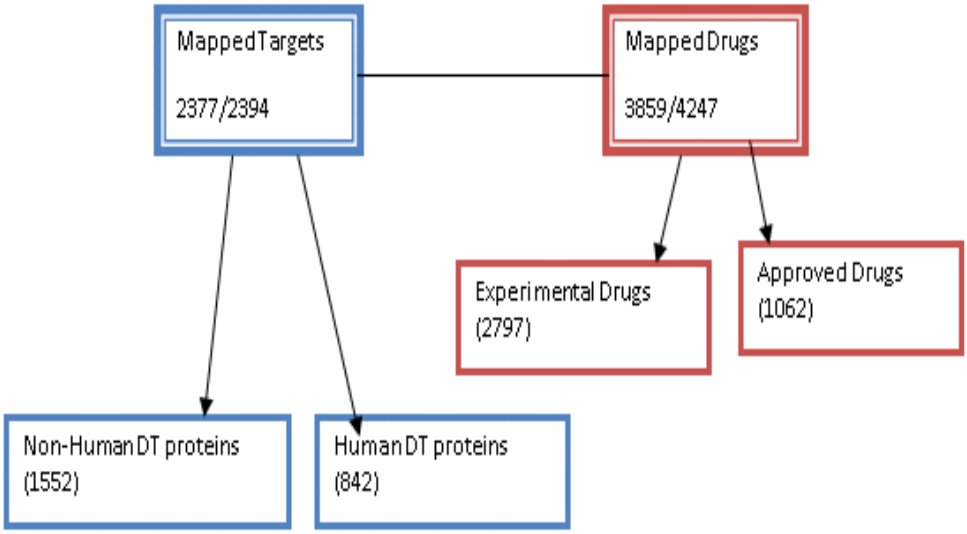


Figure 3.2: Drugbank's drug-target data mapping

3.2 Essentiality Database

3.2.1 Essentiality Data Collection

The essentiality database was created by combining data from three different sources:

1. Human essential proteins obtained from the Gandhi et al paper [24].
2. DrugBank[23] also had the essentiality information for some target proteins. The data was, therefore, queried to extract the list of proteins with their essentiality information.
3. Mouse Human orthologs essentiality data- Essential proteins data was downloaded from the MGD database in 2008. The proteins lethal to mice were termed as essential and others as non-essential. These were then mapped to their human orthologs to create a list of essential/non-essential proteins.

The records used by the above three sources were:

- Gandhi et al paper[24] - 1,206
- DrugBank[23] – 174 (human proteins)
- MGD database - 4,575

For the set of proteins identified by Uniprot IDs, Drugbank had a total of 361 essential/lethal proteins and 152 non-essential/viable proteins. In terms of human proteins, there were 32 essential/lethal versus and 142 non-essential/viable proteins (Figure 3.6). Only these human proteins were included in the essentiality database.

3.2.2 Essentiality database creation

Initially, the essentiality database was created by combining only the first two sources above, but later in the process it was found that researchers used mouse human ortholog data[2] for the essentiality related studies. To test the bias, a comparative analysis was done on the essentiality data from three sources, namely our essentiality databases, Drugbank and MGD. The two-way comparison showed the contradiction between the Drugbank and essentiality data from MGD. Ninety-one percent of proteins that were non-essential/viable according to Drugbank were termed essential/lethal in the MGD database (Figure 3.3). This shows the biased nature of MGD toward lethality.

	MGD (NE)	**MGD (E)**
DrugBank (NE)	76 (=86%)	51 (=91%)
DrugBank (E)	12(=13%)	5 (=8.9%)

Figure 3.3: Comparison of MGD and DrugBank essentiality data

However, when comparing the essentiality data of MGD with that of our essentiality database, it was found that mouse was almost 50-50% yes for both essential/lethal and non-essential/viable, which shows the uncertainty of the data (Figure 3.4).

	MGD (NE)	**MGD (E)**
Essentiality database (NE)	145(=92%)	49(=50%)
Essentiality database (E)	11(=7%)	48(=49%)

Figure 3.4: Comparison of old essentiality database and MGD

Consequently, the above comparison proves that the mouse human orthologs essentiality data is most uncertain and is more biased toward lethality/essentiality and could not be relied on for studying the essentiality of human proteins. Hence, to provide a more reliable solution, it was decided to manually integrate the data from all three sources and develop a new essentiality database. There were several criteria for curation. First, a protein would be termed as non-essential if in any of the three sources it was mentioned as non-essential/viable and termed as essential/lethal if the sources mentioned it to be lethal/essential. Second, if the information for a particular protein was not present in the other two sources, it would be considered as it was mentioned. But a protein would be dropped if it was only mentioned as lethal in MGD but absent in the other two sources since the MGD data was biased toward lethality/essentiality. It was this curated database that was further used in the study for essential proteins and drug targets. The summary flowchart of the curation process is shown in Figure 3.5.

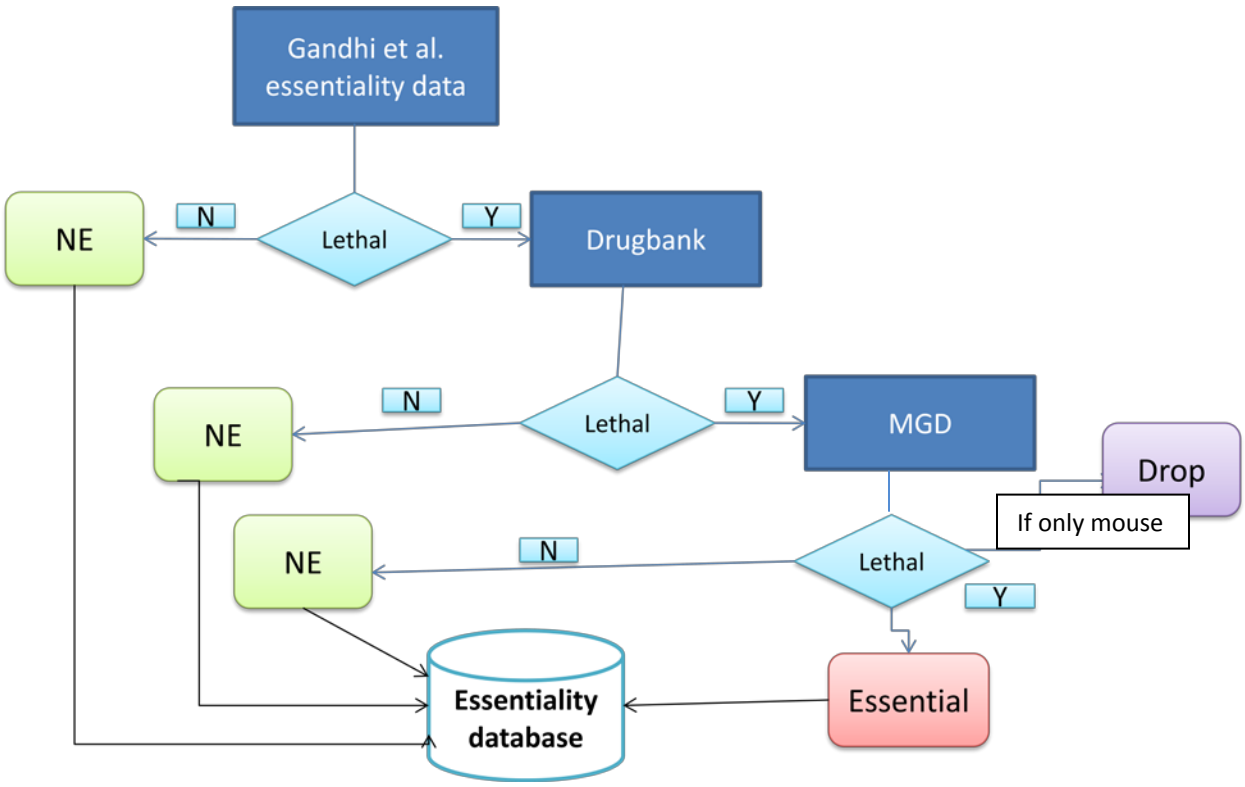


Figure 3.5: Summary flowchart of new essentiality database creation

The curated essentiality database, thus created, had a total of 3,279 proteins, of which 427 were essential/lethal and 2,852 were non-essential/viable (Figure 3.6). The essentiality database is

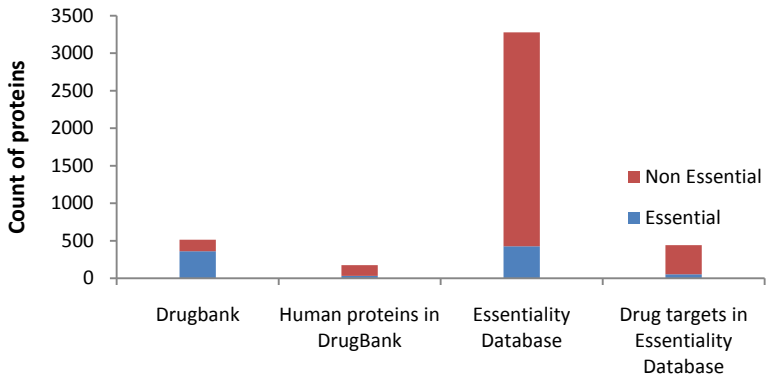


Figure 3.6: Essential and non-essential proteins in Drugbank vs. essentiality database

almost scale free (Figure 3.7) with a small representation of low degree proteins. But since it is difficult to get the data for essentiality, we could not do much about it.

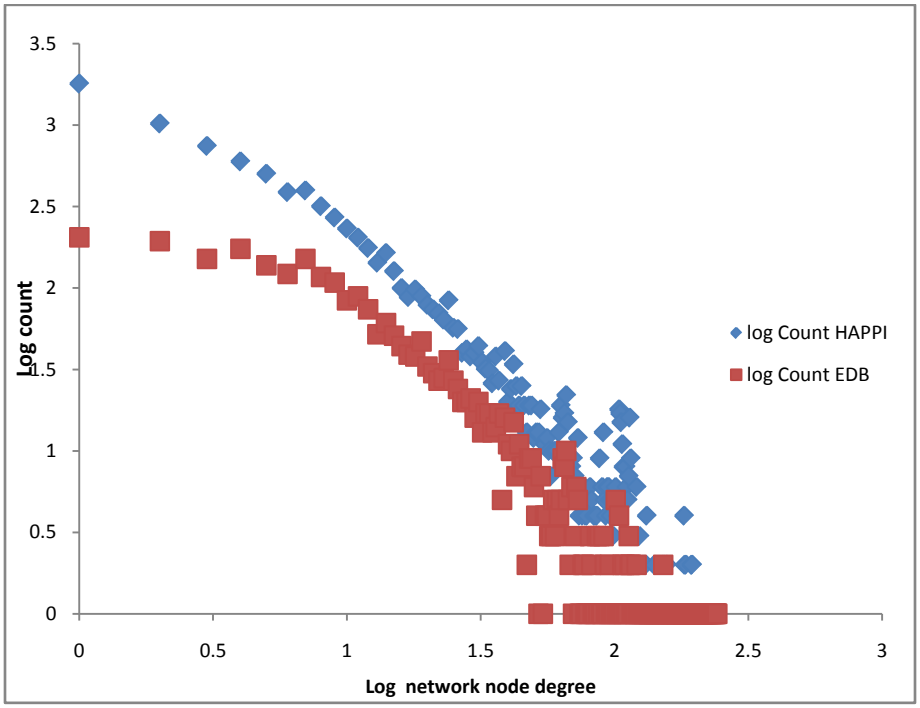


Figure 3.7: Data distribution in essentiality database

3.2.3 Essentiality Score Calculation

Drugs have multiple targets. In order to analyze the trend of target selection within drugs in terms of essentiality, the essentiality score was calculated.

$$\text{Essentiality score} = \frac{\text{No. of Essential} - \text{No. of Non} - \text{Essential}}{\text{Total target}}$$

Where:

No. of essential = the number of essential/lethal proteins targeted by a drug

No. of non-essential = the number of non-essential/viable proteins targeted by a drug

Total target = the number of targets for a particular drug

This calculation would assist in analyzing the role of essentiality in target selection or to study the existing trend in drug target proteins toward essentiality. It also would move drugs toward targeting essential/lethal proteins or non-essential/viable proteins.

3.3 Protein-Protein Interaction Data

3.3.1 Interaction Data Collection

The protein interaction data was required in order to perform the network analysis of drug targets for the hubs or non-hubs with emphasis on network degree of connectivity. The initial study was done on the subset of proteins in HAPPI database with the interaction score of above .75. In order to study different types of interactions, the data was collected from a number of other sources and the network connectivity score was calculated to make the study uniform across the different sources. The data sources considered in the study were:

- HAPPI database for different confidence ranked interactions
- STRING database for literature co-occurrences[25]
- HPRD[26]
- Pathway studio data for promoter binding
- Reactome data for metabolic pathways[27]
- Human co-expression[28]

HAPPI database (Human annotated protein-protein Interaction database). HAPPI is a comprehensive database of nearly all known protein-protein interactions. It is

an integrated source that includes interactions from SwissProt, Trembl, Online Predicted Human Interaction Database (OPHID), literature mining, etc. and focuses on human protein interactions. Based on the source of the interaction record and method of prediction, each interaction is assigned a confidence score between **0** and **1.** The interactions are then provided a confidence rank of 1-5 based on their confidence scores. The interactions from real human protein interaction experiments were provided a high score of .9 (Confidence Rank 5) and those derived from mammalian organisms were assigned a medium score. For calculating the general degree of connectivity of proteins, the interactions with the confidence score (H-score) > .75 is used. There are 9,240 interacting proteins in this range.

Confidence Rank	**Confidence score (H-score)**	**Count of interacting proteins**
Confidence Rank = 1	.1- .24	8652
Confidence Rank = 2	.25- .44	8933
Confidence Rank = 3	.45 - .74	9935
Confidence Rank = 4	.75 - .89	10056
Confidence Rank = 5	.9 - 1	5329

Figure 3.8: HAPPI database data details

STRING database[25]. This is a database of known and predicted protein-protein interactions, including both direct (physical) and indirect (functional) associations. These interactions are obtained by both experimental and predicted results and also by literature mining. The interactions obtained by mining are linked to their supporting PubMed IDs. All the proteins in the database are identified by their Protein Ensembl ID and all the interactions could be mapped to the citation information. Two proteins were said to be interacting if both had the same abstract IDs. The network degree

of connectivity of a protein was the total number of proteins it interacted with in the cut-off range. There were four cut offs applied based on the number of abstracts in which the proteins appeared together (namely 2, 3, 5 and 10).

HPRD (Human Protein reference Database) [26]. This is a data source of manually curated human protein interactions consisting of physical associations. The data could be downloaded in the tab delimited and XML formats. The raw data contained 37,107 interactions, which were then mapped to 6,927 proteins identified by uniprot ids.

Human co-expression. The data was downloaded from the supplemental data of Lees et al[28] published in 2004. The paper used 60 human microarray data sets totaling 3,924 arrays to identify pairs of genes that were reliably co-expressed based on the correlation of their expression profiles. The data was curated and a co-expression link between two genes was termed confirmed if the link was observed in more than one data set. All the supplemental data could be downloaded in MS Excel format. Only the genes with links observed in at least 3 data sources were used in the study. Once the data was downloaded, the genes were mapped to the Uniprot IDs. The degree of each protein was calculated by counting the number of co-expressed proteins for a particular protein.

Pathway Studio *(http://www.ariadnegenomics.com/products/pathway-studio/)*- This is commercial software developed by Ariadne Genomics. It is a pathway database consisting of data extracted from PubMed using Med Scan and natural language processing tools. The data from pathway studio was downloaded in MS Excel formal, which was used to download the promoter binding data. The data could not be readily mapped to the corresponding Uniprot IDs so were not used in the study.

Reactome [27]. This is a database of curated human pathways. It has the information of several pathways including metabolism, regulatory pathways, signal transduction pathways, etc. The information is curated from published literature by experts. The basic of the Reactome database is the reactions which are grouped into pathways. The reaction file could be downloaded in XML format from the reactome website. For the study, only metabolic pathways were considered. A XML parser was written in C# to parse the reactants, products, and metabolites information. Since Uniprot IDs were used as the universal identifier for the study, the metabolites would need to be mapped to their UniProt representations. But this could not be successfully done so the data was not included in the study.

3.3.2 Bayes Factor Calculation

The Bayes factor, in its simplest form, is the marginal likelihood ratio given background information. In the study, the Bayes factor was used to investigate the likelihood of a hub or a non-hub to be a drug target. The formulae for Bayes factor is

$$K = \frac{p(x|M_1)}{p(x|M_2)}.$$

With M1 and M2 being the two models and x the base of comparison.

The scale for significance of values is

- K < 3 (not significant)
- K= 3-10 (substantially significant)

For the study the formulae was arranged as

x= Drug targets (DT)

M1=hub

M2= non-hub

Thus, the formulae used in the study is

$$K = \frac{p(D.T|Hub)}{p(D.T|Non-Hub)}$$

where probability of DT given hub and non-hub was calculated by the following equation:

$$\frac{p(D.T|Hub)}{p(D.T|Non-Hub)} = \frac{(p(x = Hub/x = D.T) * p(D.T))/p(Hub)}{(p(x = non-Hub/x = D.T) * p(D.T))/p(non-hub)}$$

3.3.3 Network Connectivity Score

In order to unify the study across several other interaction datasets used in the study, a network connectivity score was calculated. The formulae for the score is

$$\textbf{Network connectivity score} = 1 - X$$

where X is the percentage of proteins at a particular network node degree range. As each database focuses on different types of interactions and thus have different network node degree for proteins, the network connectivity score would facilitate unifying the results.

3.4 Drug Side Effects Data

Since there is still no comprehensive source for the drug side effects data, it was decided to use the FDA site to pull out the data as it was the most reliable source. But the

site does not contain the side effects for all the drugs; as a result the data obtained were quite limited.

Figure 3.9 Snapshot of FDA website (http://www.fda.gov/cder/drug/DrugSafety/DrugIndex.htm)

The site provides a list of drugs along with the drugs having FDA alerts on them. Additional information about the drugs could be found by navigating to other pages. For example, the related side effects. This data is not downloadable. A parser was therefore written in python language to parse out all the related details for the drugs. The parsing of the data caused a few issues as the site had to be navigated to several pages deep to extract the information, which were in different formats. The extracted information contains the index name, use, precaution, risks, warnings, side effects, etc. and the FDA alerts, if any, related to the 572 drugs. This data then had to be mapped to the drug data present in our database. The two were mapped using the drug generic name of our drug

data to the index name, synonyms, medicine name and marketed name from the FDA side effect data. Not all drugs in our database could be mapped to the side effect data. Also it was found that one generic name could be mapped to a couple of FDA drugs identified by index names (197 generic names (drugs) could be mapped to 375 index names (FDA).) 179/197 had the side effect information.

3.4.1 Scoring Side Effects

Selecting the best criteria for scoring the side effects was a problem in itself since there is neither any ontology for side effect terms or a general list of all types of side effects. Thus, it was decided to create a list from the data itself. The warnings, side effects and risks columns were manually searched to pull out terms describing side effects. These terms were given a score of 1-5 with 5 being the deadliest side effect. These scores were further used to score the drugs (described later in drug scoring) in order to develop a unified standard for drugs. FDA alerts were put on the list initially with the idea that the drugs having the alerts should have maximum side effects, but because the frequency of the alerts was high, the idea was reconsidered. On checking, it was found that the FDA alerts were basically alerting the physicians of possible effects based on dosage and drug-drug interactions, etc. So the FDA alerts were pulled out of the list. The final list of terms and their assigned scores are as follows:

Symptom	Score	Symptom	Score
Fetal death	**5**	painful swelling	3
life-threatening	5	Abnormal vision	3
life threatening	5	Blood clot formation	3
death	5	Dehydration	3
Birth defects	4	High blood sugar	3
Bleeding	4	Increased blood sugar	3
Blood in stool	4	Pneumonia	3

Blood in urine	4	Severe diarrhea	3
Possible tumor growth	4	Shortness of breath	3
Stroke	4	Slow heart rate	3
Symptoms of congestive heart failure	4	Vomiting	3
can cause significant nerve damage	4	abnormal movements	3
cause leukemia	4	anemia	3
damage	4	blurred vision	3
depression	4	decrease in bone marrow production	3
failure blood clot	4	developed liver problems	3
may cause death	4	diarrhea	3
may worsen	4	hallucinations	3
possible development of lymphom	4	hot flashes	3
risk of serious liver injury and even death	4	hypoglycemia	3
seizures	4	leukopenia	3
serious allergic reaction	4	low blood pressure	3
serious breathing problems	4	low platelet count	3
serious condition	4	low red blood cell count	3
serious liver problems	4	myelosuppression	3
serious side effects	4	weaken your body's immune system	3
serious ulcers	4	pain	2
severe allergic reaction	4	rash	2
side effects can be severe	4	tremors	2
worsening blood disorder	4	nervousness	2
worsening of psoriasis	4	numbness	2
risk of bronchospasm	3	underactive thyroid	2
nose bleeds	3	urgent bowel movements	2
Confusion	2	redness of the eye	1
Cramps	2	loss of appetite	1
Fever	2	runny nose	1
Oily discharge	2	sleepiness	1
Swelling	2	sore throat	1
Vision problems	2	tiredness	1
burning	2	trouble sleeping	1
dizziness	2	weakness	1
fast heartbeat	2	Cold-like symptoms	1
flushing	2	loss of sleep	1
hair loss	2	red eyes	1

impotence	2	Anxiety	1
infection	2	Breast tenderness	1
joint problems	2	Cough	1
mild to moderate	2	Excessive sweating	1
mouth sores	2	Fatigue	1
Nausea	2	Gas	1
Headache	1	dry eye	1
Indigestion	1	headache	1
Menstrual changes	1	itching	1
Sweating	1	restless	1
chills	1	well tolerated	0
constipation	1	Not reported	0

Figure 3.10 Side effects scoring scheme

The scoring was subjective based on personal understanding of the seriousness of the term, but there was no other proven scoring method available.

3.4.2 Drug Score Calculation

The drugs were to be given an effect-score based on the side effects terms. For this, the terms were searched in the warnings, side effects, and FDA alert annotations for the drugs. Based on terms found in the search, the drugs were given a score. As a result, a drug(i) would have several effect scores (ISE) (Figure 3.11), as more than one term could be found in the description of side effects of drugs. Figure 3.12 shows the drug score calculation in context with the example used in Figure 3.11.

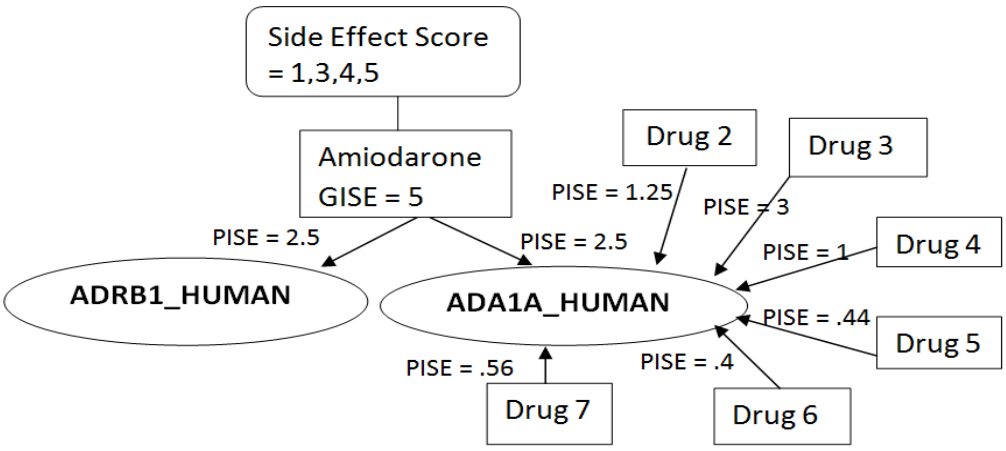


Figure 3.11: Figure showing the concept used to calculate the drug and target scores

Notation	Definition	Formulae	Example
GISE	Maximum side effect index (ISE) for drug	Max. ISE of Drug(i)	5
PISE	Partial (ISE for target	GISE(Drug(i))/number of targets	5/2 = 2.5
TISE	Total of all partial ISE of target	PISE1(T(i))+ PISE2(T(i))….	(2.5+1.25+3+1+.44 +.4+.56) = 9.15
MISE	Max. of all partial ISE of target	Maximum PISE	3
WISE	Weighted ISE of target	(PISE1*PISE1/TISE)+ (PISE2*PISE2/TISE)……	(((2.5*2.5)/9.15)+ ((1.25*1.25)/9.15) +… = 2.01

Figure 3.12: Drug and target score calculation

So the the maximum of all ISE scores for a drug(i) was taken as the drug score and termed as GISE of a drug(i) (Figures 3.11 and 3.12). This was used as a unified standard for drugs based on the side effects. The GISE ranged from 1-5 with 5 being given to the drug that had the worst life threating side effect .

$$\mathbf{GISE\,(Di) = Maximum\ of\ ISE\,(Drug(i))}$$

Since each drug targets more than one protein, the drug side effects should be the result of all the proteins it is targeting. Therefore, the drug score (GISE) was divided among all its targets (Figure 3.10). This was known as the partial probability of protein in sharing the side effect, termed PISE of target (Ti). It was calculated by the formulae (Figure 3.12)

$$\mathbf{PISE\ (Ti)} = \frac{\mathrm{GISE(Drug(i))}}{\mathrm{number\ of\ Targets}}$$

3.4.3 Target Score Calculation

The target scores were calculated in order to rank the targets based on the side effect scores. They were based on the scores provided by the drugs to target proteins. Different types of scores based on PISE values were calculated to analyze the results. Following is the formulae for target score calculation, as well as examples in Figure 3.12 of values being calculated using the Figure 3.11.

Total side effect index for target (TISE)

$$\mathbf{TISE\ (Ti)} = \mathrm{PISE\ 1(Ti)} + \mathrm{PISE2(Ti)} \ldots \ldots$$

Average side effect index for target (ATISE)

$$\mathbf{ATISE\ (Ti)} = \frac{\mathrm{TISE\ (Ti)}}{\mathrm{number\ of\ targeting\ drugs}}$$

Maximum side effect index for target (MISE)

$$\mathbf{MISE\ (Ti)} = \mathrm{maximum\ of\ PISE\ (Ti)}$$

Weighted side effect index for target (WISE)

$$\mathbf{WISE(Ti)} = \mathrm{PISE\ 1(Ti)} * \frac{\mathrm{PISE\ 1(Ti)}}{\mathrm{Total\ PISE(Ti)}} + \mathrm{PISE\ 2(Ti)} * \frac{\mathrm{PISE\ 2(Ti)}}{\mathrm{Total\ PISE(Ti)}} + \cdots$$

These scores were used to characterize the already targeted proteins that would be used in calculating the Positive Predictive Value (PPV) for measuring the true goodness of the predicted potential targets. The target score ranged from 1-5 with 1 causing the least and 5 causing the worst life threatening side effects.

3.5 Positive Predictive Value (PPV)

The PPV is the measure that reflects the probability that a positive test reflects the underlying condition being tested for. It is calculated by the following formulae:

$$\textbf{Positive Predictive Value (PPV)} = \frac{TP}{TP + FP}$$

Where

TP = proteins which ***fall within the particular network connectivity score*** and have ***target score*** **<= *2*** (less severe side-effects)

FP = proteins which ***fall within the particular network connectivity score*** but has the ***target score*** **>= *4*** (life threatening side-effects)

The PPV was calculated in order to test the true goodness of the predicted targets of the study. An example of a PPV calculation for a connectivity score of .079 - .72 for true goodness (low side effect) is as follows (values from Figure 3.13):

$$\textbf{PPV} = \frac{43}{43+22} = \textbf{.66}$$

Range(total proteins)	**.079 - .72 (75)**		**.78 - .94(32)**		**.95 - 1(3)**	
Side effect score	<= 2	>=4	<= 2	>=4	<= 2	>=4
MISE	43	22	15	12	3	0
WISE	44	19	15	9	3	0

Figure 3.13 Example table for PPV calculation

3.6 Challenges in Data Collection

The biggest challenge we faced in data collection was the mapping of the different source IDs to Uniprot IDs (ID used for the study). Some of the challenges faced were:

- We had to drop the analysis related to metabolic pathways and promoter binding because the enzymes comprised a significant amount of data that could not be mapped to Uniprot IDs.
- The mapping of the side effect data with the drugs causing them was another problem. The index names (FDA site representation for drugs) are different from the generic names (DrugBank representation for drugs). Also there were many relationships between index and generic names.
- The inconsistency in the format of the data at the FDA site caused severe problems for parsing the data.

- Due to the unavailability of side effects for all drugs, we had very little data for the analysis of ‘good targets.’

CHAPTER 4

RESULTS AND CONCLUSION

The previous chapter gave a detailed description of the biomolecular network data sets and their data sources used in the study. It also described the statistical network analysis method needed to analyze certain network properties such as network node degrees and essentiality on the datasets. When analyzing the drug target proteins for the network properties in combination with the side-effects of drugs, we reached the following conclusions:.

Results

- Essential proteins tend not to be high degree hubs (Results 4.1).
- In current drug targets, protein interaction network hubs appear to be preferentially used (Results 4.2.1).
- The preferential use of a network connectivity score by current drug targets depends on the choice of the interaction database (Result 4.2.4).
- Functional interactions with higher network node degree and physical interactions with lower to medium node degree could prove to be better drug targets (Result 4.2.4, discussed in a, c, d).
- Predicting candidate drug targets does not necessarily qualify them to be ‘good targets.’
- Essentiality, in itself, does not seem to be a predictor of drug targets (Result 4.3).
- Current drug targets seem to target more non-essential/viable proteins (Result 4.4).

The chapter provides more detailed explanation of the above results.

Results Description

4.1 *<u>High Degree Hubs Do Not Tend To Be Essential.</u>*

$$\mathrm{P(E|H\ in\ range)} = \frac{p(H|E\ in\ range\ )*p(E)}{p(H\ in\ range\ )}$$

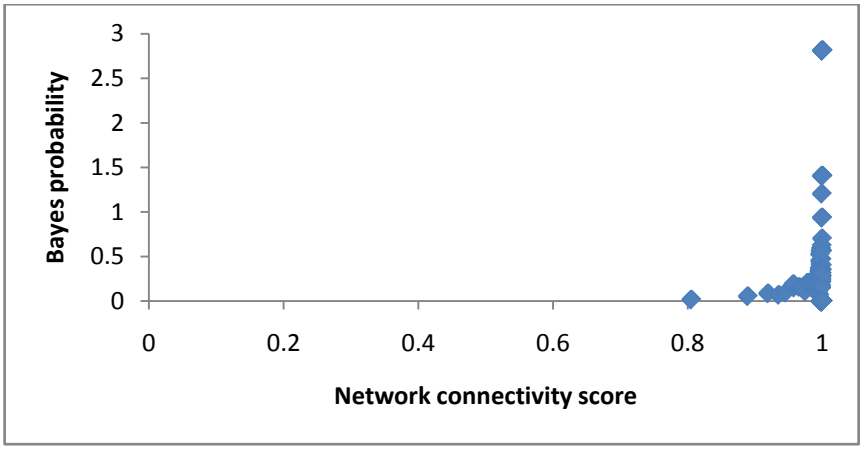


Figure 4.1Bayes probability of hubs to be essential

In Figure 4.1, we see that the Bayes probability of a hub to be essential/lethal (E) is quite low (below 1.5). If the hubs tend to be essential, Bayes probability for higher node degree should be a significant value, which is not the case. In the graph we also see that the concentration is only toward the higher connectivity score with no points toward the lower side, which indicates that the essentiality of a protein actually does not depend on the network node degree.

4.2 *<u>Hubs As Drug Targets</u>*

4.2.1 *In current drug targets, protein interaction network hubs appear to be preferentially used in retrospect.*

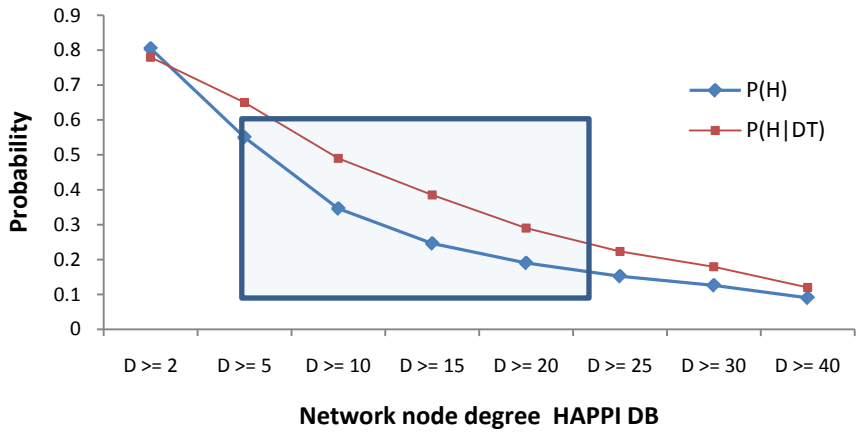


Figure 4.2: Probability of hubs at varying hub degree

In the center shaded region of Figure 4.2, we see a differential probability density between hubs and drug target as hubs. The figure shows that when the hub is defined as indistinguishable from regular proteins (at low end where hub D >=2) or very high hub (D >=40), there exists no difference between the probabilities of hubs and hubs given drug targets [p (DT|H)]. However, this difference increases for medium degree hubs and is the maximum for hub node degree 10 – 20.

From Figure 4.2, we conclude that medium network node degree proteins are more probable to be drug targets. It also promotes consideration of the topology of drug targets and node degree.

4.2.2 Preferential Selection of Hub vs. Non-hub Proteins as Drug Targets

Bayes Factor ratio *[P (DT|H) / P(DT|NH)]* was calculated to show the preferential use of drug targets in hubs vs. non-hubs. To take the arbitrary definition of "hub" and "non-hub" out of the picture and show the significance of

the ratio, we performed our analysis by varying the definitions of hubs and non-hubs network node degrees.

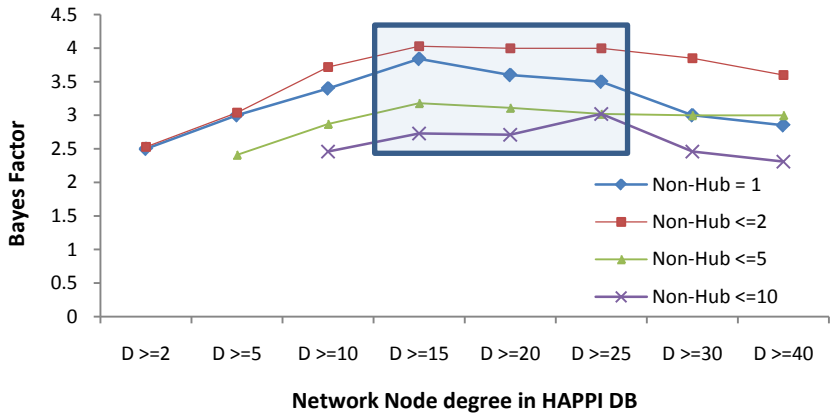


Figure 4.3: Preferential use of drug targets in hubs vs. non-hubs

The peak in Figure 4.3 seems to point out that neither light hubs nor heavy hubs are good existing "drug targets." However the non-hubs definition of network node degree <=2 and hub definition of network node degree in the range of 15-25 has the maximum Bayes factor.

This gives an indication that the proteins within the network node degree range of 15-25 would be promising candidates for drug targets no matter how we define the hubs and non-hubs.

4.2.3 Case Study on Cancer Proteins

The analysis result of medium degree proteins as probable drug targets was tested on cancer proteins (see methods chapter). The subset of 66 of 382 proteins passed the network topology filter of 15-25. Twenty-three of 66 proteins were existing targets of cancer drugs (Figure 4.4). On searching the PubMed for the remaining 43 proteins, it

was found that 18 were reported as promising targets, with 4 as candidate targets (see appendix, Figure 4.21).

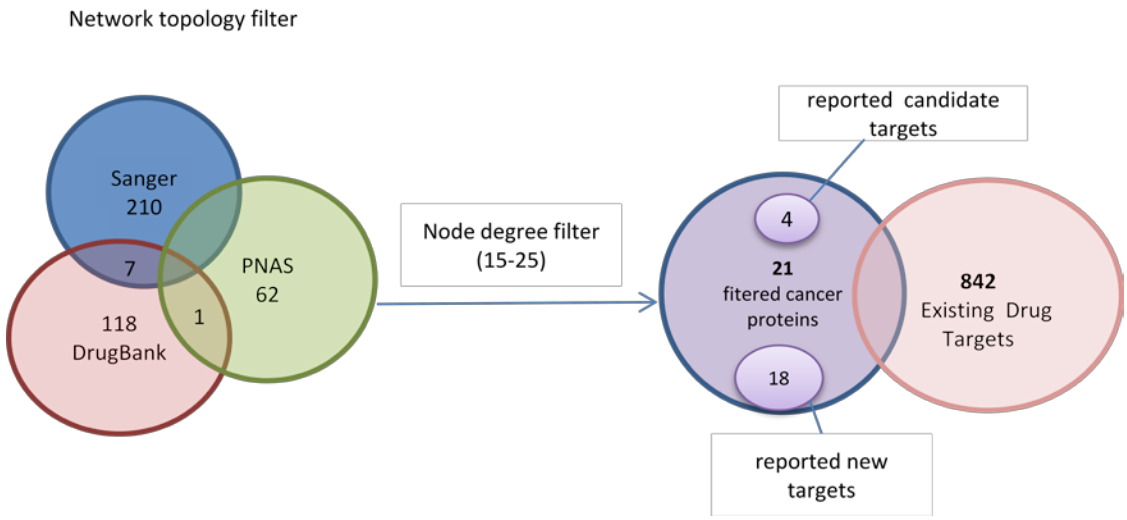


Figure 4.4 Statistics of Cancer case study

Thus, 68% [(23+18+4)/66] of the medium degree hub proteins were either current targets or indicated to be promising targets for drugs. But the question is whether these targets are really 'good targets' that would cause minimal side effects when targeted by drugs.

4.2.4 *Probability of Hubs, Derived from Different Data Sources, to be Drug Targets*

To analyze the results on different types of interaction datasets, the study was extended to other databases (see method). In order to ensure uniformity across all datasets, a Network Connectivity Score was calculated (see method section). We then calculated the Bayes factor at varying definitions of non-hubs and the Network connectivity scores.

Overview of all the data sources-The non-hub definition of network node degree <= 2 consistently gave the best Bayes factor (see appendix Figure 2-4), so it was considered to be the base line for non-hubs further in the study. Of all rated

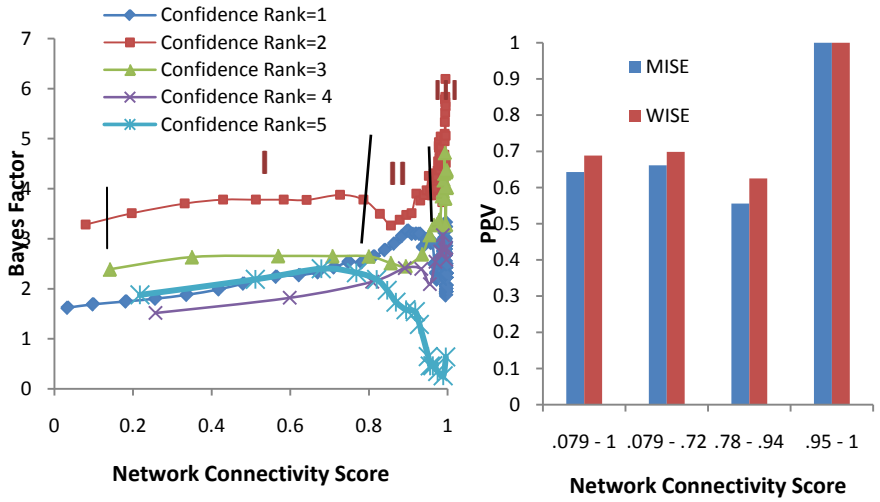
interactions in HAPPI, the confidence ranking of 2 and >= 2 gave the best Bayes factor (Figures 4.5(a) and 4.6(a), respectively). In literature co-citations, the best Bayes factor was for cut-off 2 (Figure 4.7(a)). However, the Bayes factor drops as the cut-off for citations becomes stringent (cut-off 5, 10), indicating the inconsideration toward the stability of interactions in selecting drug target proteins. HPRD interactions gave a low Bayes factor (highest value of 2.5 (Figure 4.8(a)). The lowest Bayes factor (less than 1.5) was for co-expression (see appendix Figure 4), indicating that considering co-expression as a characteristic while selecting drug target proteins would not be a good idea.

Detailed Analysis of data with best Bayes Factor

a) HAPPI –

i. Confidence Rank =2

Figure 4.5(a) is the Bayes factor graph for the HAPPI ranked interactions (confidence rank 1-5). The true measure of the candidate drug targets for each of the sections I, II, and III is shown in the Positive Predictive Value (PPV) graph (Figure 4.5(b)). Figure 4.5(c) shows the data values of the PPV graph and Figure 4.5(d) uses the cancer proteins to predict 'good targets.'



(a) *(b)*

Range(total proteins)	.079 - 1 (113)		.079 - .72 (75)		.78 - .94(32)		.95 - 1(3)	
Side effect score	<= 2	>=4	<= 2	>=4	<= 2	>=4	<= 2	>=4
MISE	63	35	43	22	15	12	3	0
WISE	64	29	44	19	15	9	3	0

(c)

Figure 4.5 (a) shows Bayes factor graph for HAPPI confidence rank (1-5); (b) (c) shows the PPV and their data values for the ranges

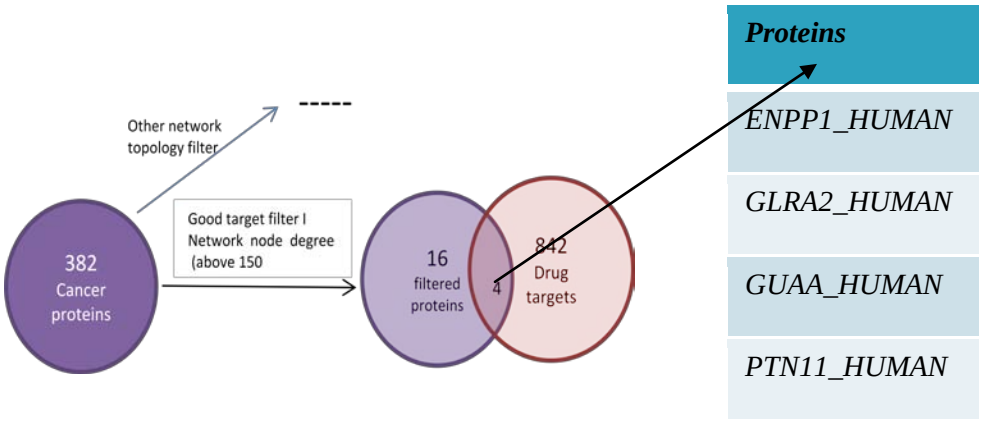


Figure 4.5(d) Case study of 'good targets'

Figure 4.5(a) shows the best Bayes factor (value above 3) is given by the interactions with the confidence ranking of 2 (low quality not

experimentally verified interactions). The graph shows an increase in Bayes factor for the score range 0.1- 0.7 (part I of Figure 4.5(a)), but then decreases and increases again to give a significant value for the score of 0.95-1 (part II of Figure 4.5(a)). The ranked 2 proteins also give an overall 70% PPV with a connectivity score of .95-1 (network node degree >= 150, see appendix) giving a PPV of 1 (100%) (Figure 4.5(b). However, we see a low Bayes factor for interactions with highest ranking (confidence rank of 5) in HAPPI, which are verified physical interactions. In this case, the Bayes factor value increases for the score range 0.2-0.6, but then declines for high network node degree proteins.

We conclude from Figure 4.5 (a) that among all ranked proteins in HAPPI, those ranked 2 would be the best candidates for drug targets. However, lower PPV for the connectivity score ranges other than .95 – 1 indicates that predicting candidate drug targets does not essentially qualify them to be 'good targets' causing minimal side effects. When applying the connectivity range with the best PPV to cancer proteins (Figure 4.5 (d)), it was found that out of 16 proteins which passed the filter, 4 were existing targets. Of these 4 proteins (Figure 4.5(d)), 2 were existing targets of approved drugs. On analyzing the side effects of the drugs targeting these proteins, it was found they were well tolerated.

The low quality interactions giving the best Bayes factor indicates the possibility of **functional interactions being preferred over physical interactions in current target selection.**

Drug target	Targeting Drug (number of targets)	WISE	Drug Symptoms	Therapeutic area
AT1A1_HUMAN (Degree =483)	Pantoprazole(1)	1	Headache, well tolerated	Short-term treatment of erosive esophagitis.
5HT2C_HUMAN (Degree =448)	a. Mirtazapine(4) b. Quetiapine(4) c. Ziprasidone(9)	1.03	a.Fever, Gas ,dizziness, Stroke, depression b.confusion, headache, fever, death,dizziness c.constipation, dizziness, restlessness, diarrhea, rash	a.Depressive disorder. b.schizophrenia and acute manic episodes c. schizophrenia
5HT2B_HUMAN (Degree= 192)	a. Quetiapine(4) b. Eletriptan(7)	1.06	Depression, dizziness, rash, weakness, depression, life-threatening condition	migraine with or without aura in adults.

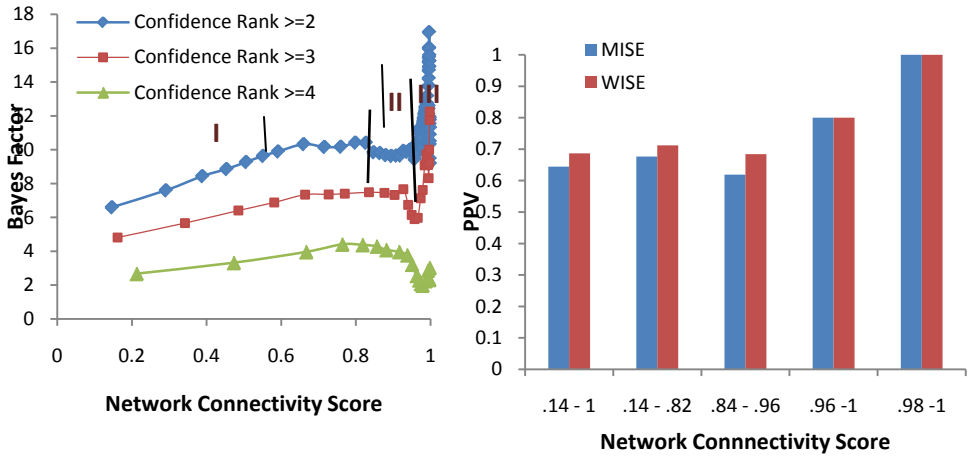
Figure 4.5(e) Table for good drug target from HAPPI rank 2 where PPV = 1

The drugs such as Ziprasidone, Quetiapine, and Eletriptan may cause serious side-effects but only if treated with other depression medicines or for dementia.

ii. Confidence Rank Above 2, 3 and 4

Figure 4.6 (a) shows the interactions from HAPPI with the confidence ranking of 2 and above, and the Positive Predictive Value (PPV) for different sections (I,II,III,IV) has been shown in Figure 4.6 (b). The data values for PPV of all ranges are shown in Figure 4.6(c). Among all confidence rankings, 2 and above have the highest and the most significant Bayes factor. While interactions with confidence ranking of 3 also gave a good Bayes factor, they were not considered for further analysis as they

have potentially transient interactions and may not be stable. So interactions ranked 2 and above were further analyzed.

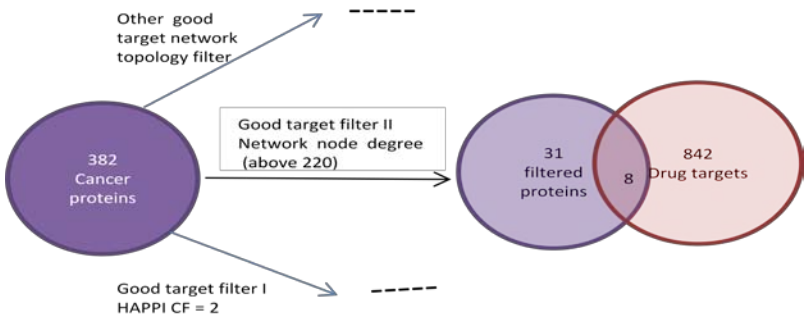


(a) *(b)*

Range(total proteins)	.14 - 1 (119)		.14 - .82 (81)		.84 - .96(25)		.96 - 1(5)		.98 - 1(2)	
Side effect score	<= 2	>=4	<= 2	>=4	<= 2	>=4	<= 2	>=4	<= 2	>=4
MISE	67	37	43	22	13	8	4	1	2	0
WISE	68	31	46	22	13	6	4	1	2	0

(c)

Figure 4.6 (a) Bayes factor graph for confidence rank above (2, 3, 4); (b) (c) PPV and data values for HAPPI (confidence rank 2 and above); (d) Case study 'good targets'



(d)

We see in Figure 4.6(a) that for confidence rank >= 2 there is an increase in Bayes factor for the connectivity score range of 0.2-0.8, which decreases for a small range before increasing again (connectivity score range .96-1 (node degree above 220, see appendix), thus giving a high significant Bayes factor value. We see again that the PPV values are different for all ranges. Figure 4.6(b) shows the overall PPV for rank 2 and above to be almost 70% with 100% for the connectivity score of .98 – 1 (network node degree >= 320, see appendix). The interactions with confidence rank of 4 and above give the Bayes factor in the range 2-4, with only a small range above 3.

High Bayes factor by interactions ranked 2 and above and lower by ranked 4 and above gives a clear indication that **functional interactions are being considered over stable physical interactions in drug target selection.** This also shows that the Bayes factor value gradually starts decreasing as the interactions become more and more significant. The different PPV values in all connectivity ranges further strengthens our belief that **predicting candidate drug targets does not essentially qualify them to be 'good targets.'** When applying the connectivity range filter to cancer proteins (Figure 4.6 (d)), we found that out of 31 proteins which passed the filter, 8 were existing targets.

Drug target	Targeting Drug (number of targets)	WISE	Drug Symptoms	Therapeutic area
5HT2C_HUMAN (Degree =595)	a. Mirtazapine(4) b. Quetiapine(4) c. Ziprasidone(9)	1.03	a.Fever, Gas ,dizziness, Stroke, depression b.confusion, headache, fever, death,dizziness c.constipation, dizziness, restlessness, diarrhea, rash	a.Depressive disorder. b.schizophrenia and acute manic episodes c. schizophrenia
AT1A1_HUMAN (Degree =483)	Pantoprazole(1)	1	Headache, well tolerated	Short-term treatment of erosive esophagitis.

Figure 4.6(e) good drug target from HAPPI Confidence rank >= 2 with PPV = 1

b) STRING (Literature co-occurrence)

Figure 4.7 (a) shows the Bayes factor graph for literature co-occurrence with the cutoff of 2, 3, 5, and 10. The Positive Predictive Value (PPV) for different sections (I, II, III) has been shown in Figure 4.7(b) with its data values in Figure 4.7(c).

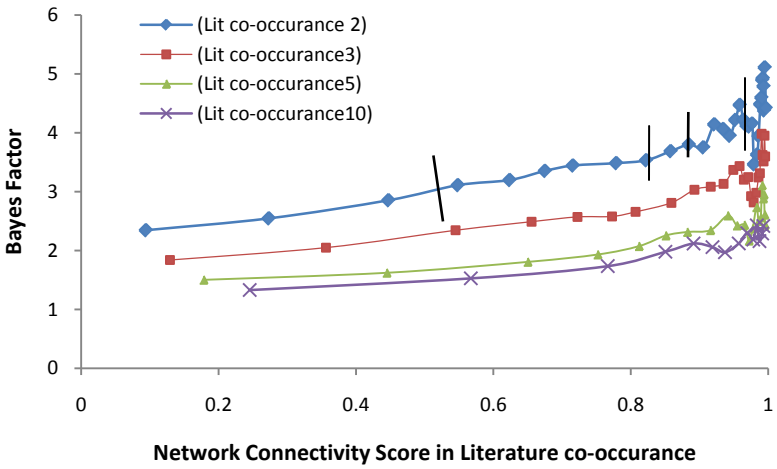
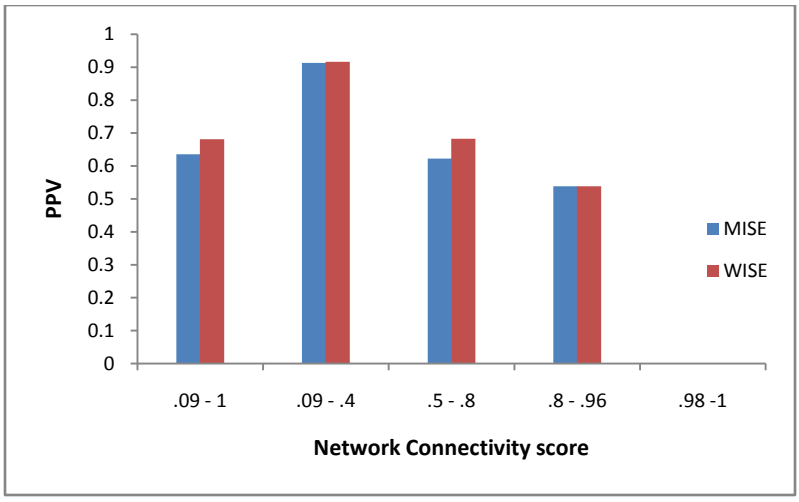


Figure 4.7(a)



(b)

Range(total proteins)	.09 - 1 (110)		.09 - .4(26)		.5- .8 (54)		.8 - .96(14)		.98 -1(2)
Side effect Score	<= 2	>=4	<= 2	>=4	<= 2	>=4	<= 2	>=4	<= 2
MISE	61	35	21	2	28	17	7	6	0
WISE	62	29	22	2	28	13	7	6	0

(c)

Figure 4.7(b) and (c) shows the PPV for Co-occurrence cut-off 2 and their data values, respectively

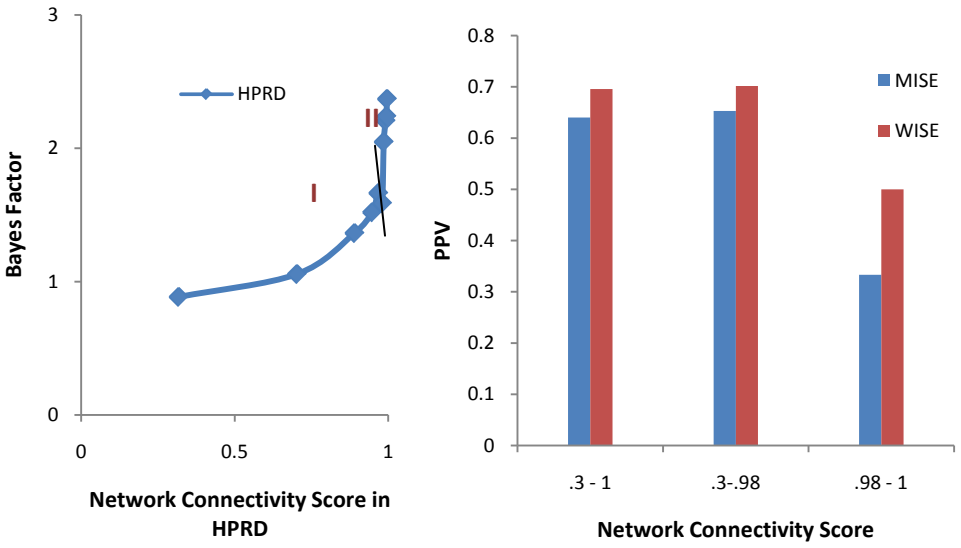
In Figure 4.7(a), we see that the co-occurrence cutoff 2 gives the highest Bayes factor value. However, it decreases and also falls below any significant value for cutoff 10 (stable interactions). Co-occurrence cutoff 2 interactions are the least cited and most probably the functional interactions. Here again we see different PPV for different connectivity ranges. We get an overall PPV of almost 70% (Figure 4.7(b)) with the connectivity score of .5-.8 giving better PPV than the higher connectivity score.

The high Bayes factor for co-occurrence cutoff 2 further proves our point that **stable physical interaction is not the current criteria of selecting drug target proteins.** We also conclude from studying all the sources showing different PPV

for different connectivity scores above that **selecting drug target does not necessarily make it a 'good target.'** The high PPV for medium connectivity score with a low Bayes factor gives an indication that lower-to-medium degree proteins could be better targets causing minimal side effects if literature co-occurrences or single interaction data sources are being considered in target selection.

c) HPRD

Figure 4.8 (a) shows the Bayes Factor graph for HPRD; the Positive Predictive Value (PPV) for sections I and II is shown in Figure 4.8 (b).



(a) *(b)*

Range(total proteins)	.3 - 1 (87)		.3-.98(84)		.98-1 (3)	
Side effect score	<= 2	>=4	<= 2	>=4	<= 2	>=4
MISE	48	27	47	25	1	2
WISE	48	21	47	20	1	1

(c)

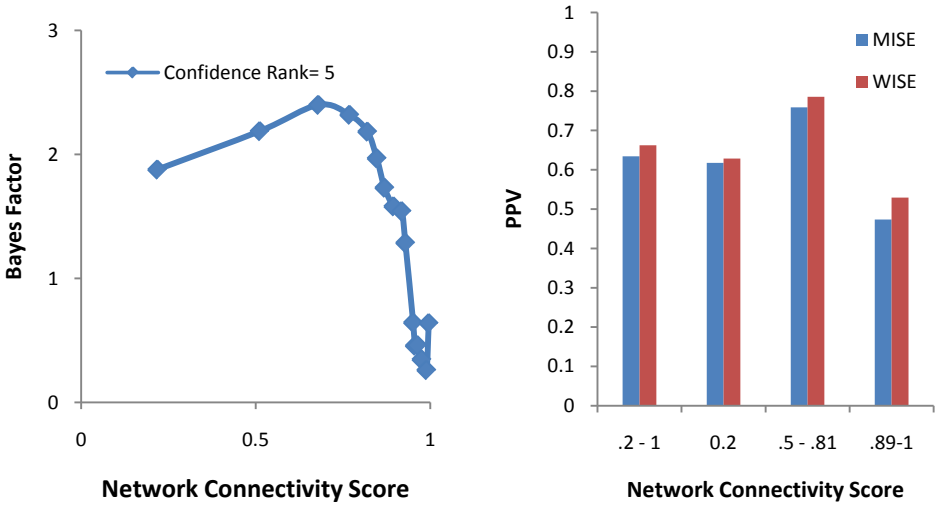
Figure 4.8 (a) shows the Bayes factor; (b), (c) shows PPV graphs and data values for the score ranges respectively.

We can see from Figure 4.8(a) that an HPRD interaction does not give a significant Bayes factor but gives an overall 70% PPV (Figure 4.8(b)). The graph also shows a higher PPV for a lower-to-medium connectivity score with a lower PPV for a higher connectivity score.

The low Bayes factor for HPRD interactions further proves our theory and we conclude that **stable physical interactions are not preferred candidates for drug targets.** The low PPV for high connectivity score is another indication that low-to-medium network node degree proteins would be a better source for 'good targets' when considering independent sources for interactions.

d) High-Ranked Interactions in HAPPI

The highly-ranked interactions in HAPPI with the confidence ranking of 5 are physical interactions which have been validated across several other sources.



(a) *(b)*

Figure 4.9(a) and (b) shows the Bayes factor and PPV graphs for HAPPI ranked 5 interactions, respectively.

From Figure 4.9(a), we see that the high-rated interactions do not give significant Bayes factor, which again shows that specific physical interactions are not considered in the selection of drug target proteins. However, medium degree hubs could be good targets if specific interactions or physical interactions are being considered.

4.3 Current Drug Targets and Essentiality

4.3.1 Essentiality not a Good Predictor for Selecting Drug Targets.

We calculated the Bayes factor [p (DT|E) / p (DT|NE)] to find if essentiality plays a significant role in the selection of drug targets.

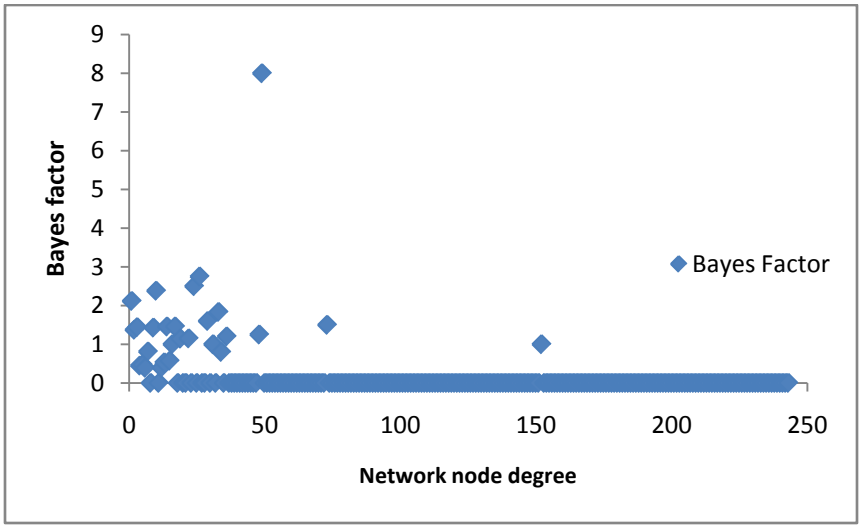


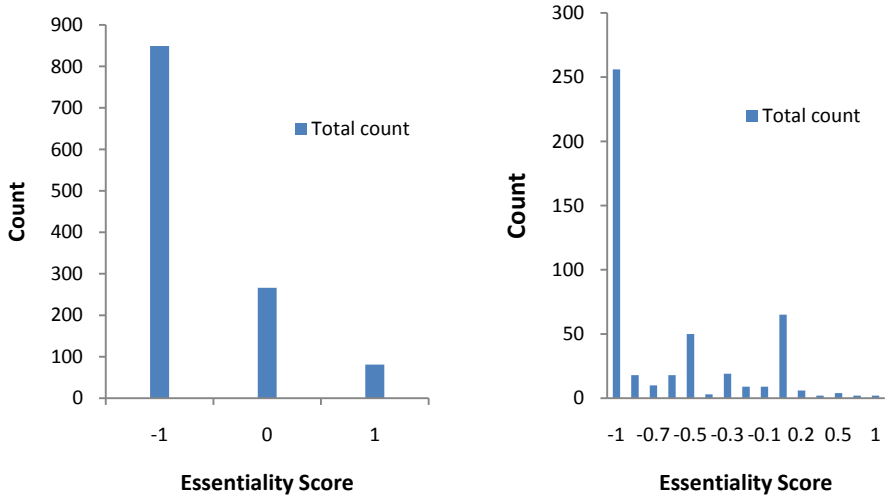
Figure 4.10: Essentiality not a predictor for drug targets

Figure 4.10 show that essentiality as a predictor gives a low Bayes factor value. We can also see one point with a high Bayes factor because in that particular range there is only one essential protein, which is also a drug target.

The low Bayes factor value is an indication that essentiality alone would not prove to be a very good predictor of drug target.

4.3.2 *Non-essential/viable Proteins as Current Drug Targets*

In order to analyze the trend in drug target proteins toward essentiality of proteins, we calculated the essentiality score (see methods) and analyzed them for drugs with single and multiple targets (Figures 4.12 (a) and (b).) We also compared them for FDA- approved and experimental drugs to see if there is a difference in trend (see appendix Figure 5).



(a) (b)

Figure 4.11 (a) Single-targeted drugs and (b) Multi-targeted drugs

Figures (a) and (b) above clearly show that non-essential/viable proteins are being targeted more than essential/lethal proteins.

By seeing the trend in approved drugs, we can conclude that non-essential/viable proteins could be better future drug targets.

4.4 *Bad Targets*

We created a list of potential 'bad targets' (Figure 4.12) based on the target scores (WISE = 5). When analyzed for their network node degree, it was

found that most of them did not fall in the range of our criteria for candidates of good drug target node degree, which is as follows:

- HAPPI confidence rank 2 – connectivity score .95-1 *(PPV = 1)*
- Literature co-occurrence 2- connectivity score .09-.4 *(PPV=.91)*

Drug targets	***Targeting Drugs(WISE)***	***Degree rank=2 (Connectivity score)***	*Lit co-occurrence 2* ***(Connectivity score)***
AOFA_HUMAN		*<.95*	*>.4*
CD52_HUMAN	*1(5)*	*<.95*	*>.4*
PRGR_HUMAN	*1(5)*	*<.95*	*>.4*
V2R_HUMAN	*1(5)*	*<.95*	*>.4*
PDE4B_HUMAN	*1(5)*	*<.95*	*>.4*
NEU2_HUMAN	*1(5)*	*<.95*	*>.4*
KCNK2_HUMAN	*1(5)*	*<.95*	*>.4*
PYRD_HUMAN	*1(5)*	*<.95*	*>.4*
ACES_HUMAN	*2(5)*	*<.95*	*>.4*
ACET_HUMAN	*8(5)*		

Figure 4.12 List of 'bad targets' and their connectivity score

4.5 CONCLUSION

- Network Biology perspective to drug discovery is necessary when selecting candidate drug targets and could bring forth a new approach to target selection.
- Selecting drug targets does not necessarily qualify them as 'good targets.'
- Functional interactions are preferred in the selection of candidate drug targets.
- Integrated sources are a better source for interactions of candidate drug targets.

- Essentiality in itself is not a very good predictor for selection of targets.
- Current drug targets tend to target non-essential proteins.

CHAPTER 5

DISCUSSION

The network biology perspective in studying candidate drug targets is necessary due to the complexity of molecular systems. To perform our study about drug selection from the network biology perspective, we used the emerging biomolecular data sets, which included drug and target, protein interaction, literature co-occurrence, drug side effects and essentiality. We used Bayes factor and Positive Predictive Values to examine the use of certain network properties, such as network node degrees and essentiality, to predict candidate drug targets. Using this method, we were able to find the best network connectivity scores for current targets; this method also will assist in predicting candidate drug targets.

We also showed the distinction between drug targets and 'good drug targets.' We were able to predict the connectivity scores for 'good targets,' which could be used to find probable drug targets causing minimum side effects. We ranked the drugs based on their side effects in an effort to standardize them. By taking into account the aggregated side effect scores of all FDA-approved drugs, we developed a metric for their target proteins. In the process, we developed the essentiality database.

The low availability of side effect data and lack of common terminology for the side effects, however, created some problems. We had to score them per our understanding, which may not have resulted in the best scores for a particular description of side effect. The essentiality database created was not a true representation of low node

degree proteins but could not be helped because of the low availability of the essentiality data.

There is a need for detailed side effect information of drugs with standard terminology. An integrated database for drugs, their side effects, drug targets and their interacting partners, along with their homolog proteins and their pathway details, would be a good source for analyzing drug targets.

APPENDIX

1. Literature mining for potential cancer Drug Targets

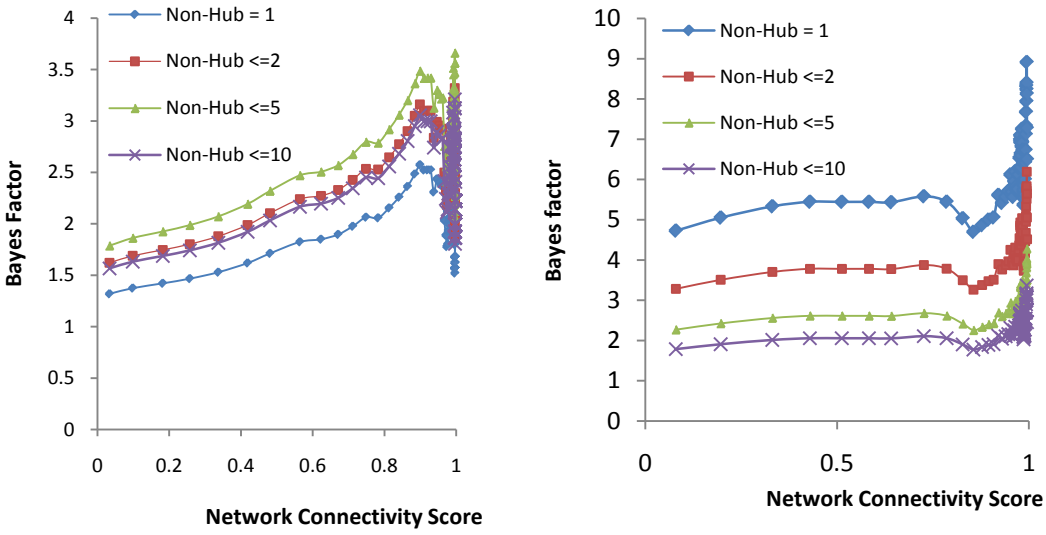
Drug target protein in filter range	Node degree	Gene name	Current drug target	Literature references
BRAF1_HUMAN	25	BRAF		use of this gene as target for an effective cancer therapy.
IMA2_HUMAN	25	KPNA2		
MET_HUMAN	25	MET	Y	"*ETV6-NTRK3—Trk-ing the primary event in human secretory breast cancer*"- role of Trk signaling in breast cancer and also suggest a target for drug development.
NTRK3_HUMAN	25	NTRK3		
TIF1A_HUMAN	25	TIF1		
TRAF4_HUMAN	25	**TRAF4**		"P300/CBP acts as a coactivator to cartilage homeoprotein-1 (Cart1), "lysine may be a common target for HAT of p300/CBP for these proteins.
TSC1_HUMAN	25	TSC1		
TSHR_HUMAN	25	TSHR	Y	
DHSB_HUMAN	24	SDHB		succinate dehydrogenase deficiency may be the cause of a subgroup of GISTs and this offers a therapeutic target for GSTs Gastrointestinal stromal tumors(GISTs)
KCY_HUMAN	24		Y	
MAF_HUMAN	24	MAF		
PDGFB_HUMAN	24	PDGFB		
PGFRA_HUMAN	24	PDGFRA		FIP1L1-PDGFRA fusion gene are excellent candidates for treatment with tyrosine kinase inhibitors even if they present with an aggressive phenotype such as AML."
TEC_HUMAN	24	TEC		
TOP1_HUMAN	24	TOP1		
BLM_HUMAN	23	BLM		
GATA1_HUMAN	23	GATA1		
INAR1_HUMAN	23	**IFNAR1**	Y	
FGFR2_HUMAN	22	FGFR2	Y	
FLT3_HUMAN	22	FLT3		FLT3/ITD and can be a therapeutic target in the treatment of AML with FLT3/ITD". Okamoto, M.Hayakawa, F.Miyata
FOXO1_HUMAN	22	FOXO1A		
HS90B_HUMAN	22	HSPCB	Y	
MLH1_HUMAN	22	MLH1		

MOES_HUMAN	22	MSN		
PA1B3_HUMAN	22	PAFAH1B3		
PRGR_HUMAN	22	PGR	Y	
PTMA_HUMAN	22	**PTMA**		PTMA expression in RMS biopsy samples might prove to be an effective diagnostic marker for this disease
TSC2_HUMAN	22	TSC2		
ADT3_HUMAN	21		Y	
FGFR3_HUMAN	21	FGFR3		
LDLR_HUMAN	21		Y	
NF1_HUMAN	21	NF1		
RBTN2_HUMAN	21	LMO2		
ALK_HUMAN	20	ALK	Y	
BCL10_HUMAN	20	BCL10		
IMDH1_HUMAN	20		Y	
MERL_HUMAN	20	NF2		“A clue to the therapy of neurofibromatosis type 2: NF2/merlin is a PAK1 inhibitor”
NPM_HUMAN	20	NPM1		NPMI mutation may represent a new target to monitor minimal residual disease in AML and a potential candidate for alternative and targeted treatments
PPAT_HUMAN	20	PPARG	Y	
ADT1_HUMAN	19		Y	
BTG1_HUMAN	19	BTG1		BTG1 could be used as a potential treatment-related biomarker for monitoring the therapy effect
FLI1_HUMAN	19	FLI1		
MTG8_HUMAN	19	CBFA2T1		RUNX1-CBFA2T1 is a promising and leukaemia-specific target for molecularly defined therapeutic approaches.
TYSY_HUMAN	19		Y	
METH_HUMAN	18		Y	
RARB_HUMAN	18			
SAHH_HUMAN	18		Y	
SOAT1_HUMAN	18		Y	
TCPE_HUMAN	18	**CCT5**		CCT5 clinically useful in identifying the subset of breast cancer patients who may or may not benefit from docetaxel treatment.
WRN_HUMAN	18	WRN		a possible therapeutic role for WRN as an anti-cancer target, and highlight the importance of WRN protein status for tumorigenesis and clinical treatments of patients
ETV6_HUMAN	17	ETV6		
IRF4_HUMAN	17	IRF4		IRF4 can be used as a molecular marker of clinical subtype in ATL.'
LOX15_HUMAN	17	**ALOX15**		
PAX3_HUMAN	17	PAX3		'PAX3 as a promising target for immunotherapy of cancer.'
TCPD_HUMAN	17	CCT4		

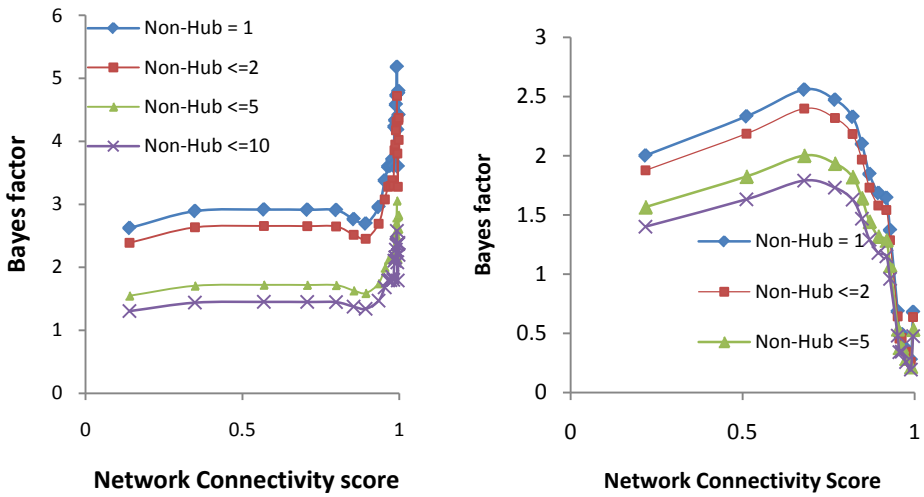
ATF1_HUMAN	16	ATF1		
CBX3_HUMAN	16	CBX3		
CO1A2_HUMAN	16	COL1A2		
FOXO3_HUMAN	16	FOXO3A		FOXO1 protein appears to be a promising target for future drug discovery and cancer therapy.
MEN1_HUMAN	16	MEN1		
PRLR_HUMAN	16		Y	
RXRB_HUMAN	16	RXRB		
ANXA1_HUMAN	15	ANXA1	Y	
FANCC_HUMAN	15	FANCC		The impact of Fanconi gene defects on drug and irradiation sensitivity renders these genes promising targets for a specific, genotype-based therapy for individual cancer patients, providing a strong rationale for clinical trials.
PIM1_HUMAN	15	PIM1	Y	
TPM3_HUMAN	15	TPM3		TM gene expression is a target of oncogene action or is an indirect consequence.

2. *<u>Bayes Factor at varying definitions of Hub and Non-Hub.</u>*

   a) *HAPPI Confidence rating 2-5 respectively (Figure 1 a-d).*



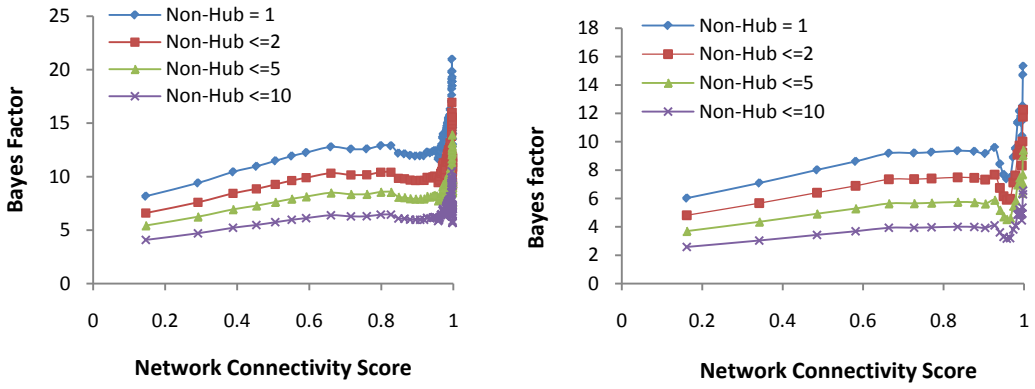
(a) (b)



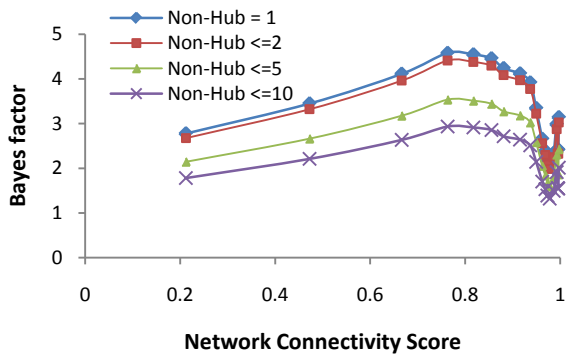
(c) (d)

Figure 1: Bayes factor graph for (a) confidence rating 2; (b) confidence rating 3; (c) confidence rating 4 (d) confidence rating 5

b) HAPPI Confidence rating above 2, 3 and 4 respectively (Figure 2 a-c).



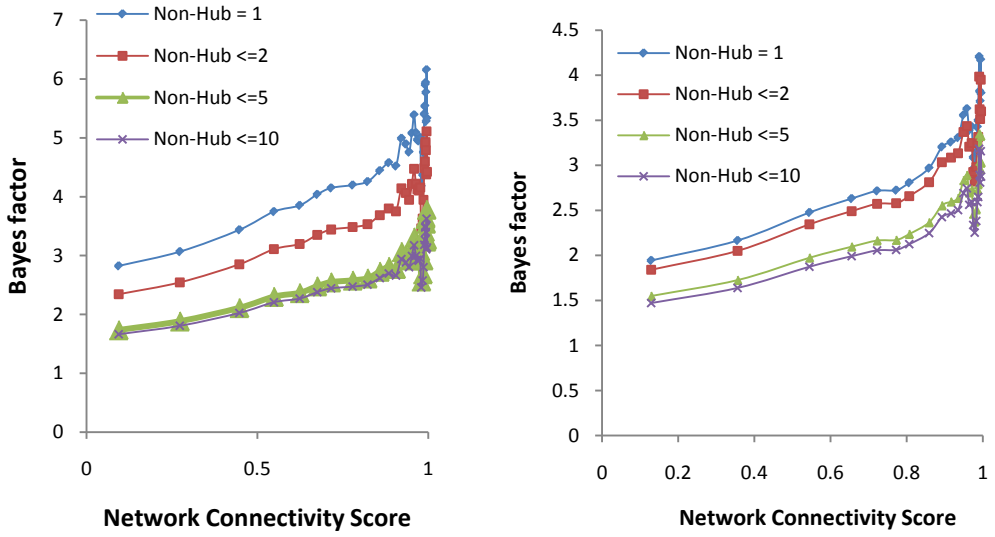
(a) (b)



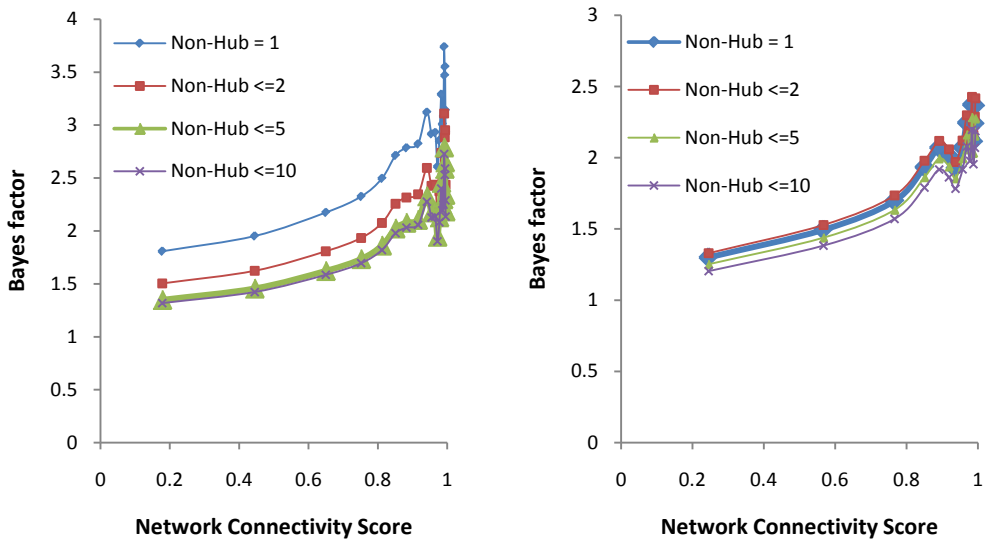
(c)

Figure 2: Bayes factor graph for (a) confidence rating above 2; (b) confidence rating above 3; (c) confidence rating above 4

c) Literature co-occurrence cut-off 1, 2, 5, 10 (Figure 3 a-d)



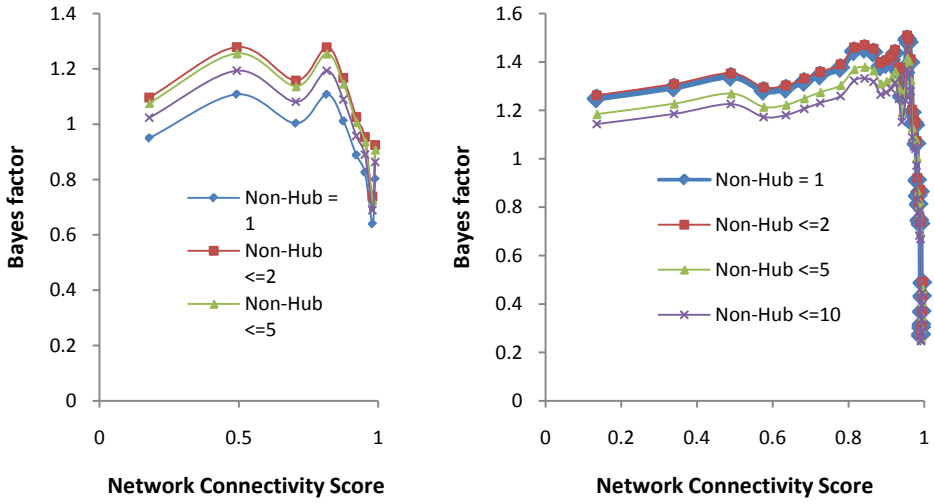
(a) (b)



(c) (d)

Figure 3: Literature co occurrence (a) cut-off 2; (b) cut-off 3; (c) cut-off 5; (d) cut-off 10

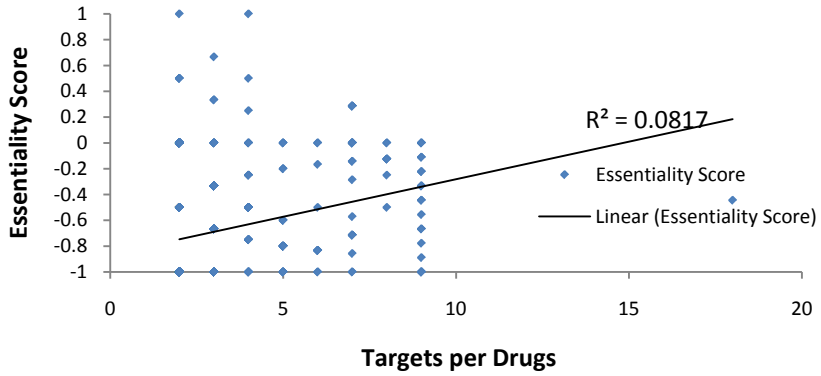
d) Co-expression (Figure 4 a-b)



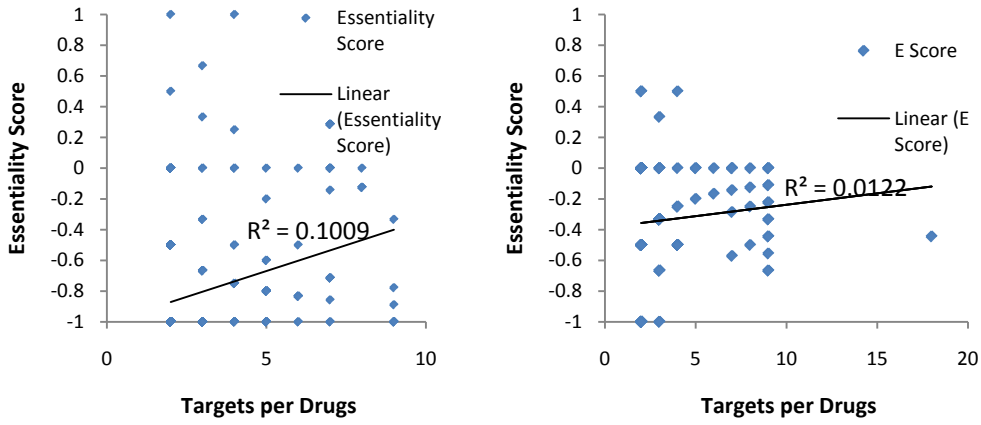
(a) (b)

Figure 4 (a) Conserved Co-expression; (b) Human co-expression

3. Co-relation between essentiality of targets and Multi targeted Drugs (Figure 5 a-c).



(a)



(b) (c)

Figure 5 (a) All Drugs; (b) Approved Drugs; (c) Experimental Drugs

REFERENCES

1. Hopkins, A.L., *Network pharmacology.* Nat Biotechnol, 2007. **25**(10): p. 1110-1.
2. Yildirim, M.A., et al., *Drug-target network.* Nat Biotechnol, 2007. **25**(10): p. 1119-26.
3. Frantz, S., *Drug discovery: playing dirty.* Nature, 2005. **437**(7061): p. 942-3.
4. Roth, B.L., D.J. Sheffler, and W.K. Kroeze, *Magic shotguns versus magic bullets: selectively non-selective drugs for mood disorders and schizophrenia.* Nat Rev Drug Discov, 2004. **3**(4): p. 353-9.
5. Hopkins, A.L. and C.R. Groom, *The druggable genome.* Nat Rev Drug Discov, 2002. **1**(9): p. 727-30.
6. Jeong, H., et al., *Lethality and centrality in protein networks.* Nature, 2001. **411**(6833): p. 41-2.
7. Han, J.D., et al., *Evidence for dynamically organized modularity in the yeast protein-protein interaction network.* Nature, 2004. **430**(6995): p. 88-93.
8. Csermely, P., V. Agoston, and S. Pongor, *The efficiency of multi-target drugs: the network approach might help drug design.* Trends Pharmacol Sci, 2005. **26**(4): p. 178-82.
9. Kunkel, E.J., *Systems biology in drug discovery.* Conf Proc IEEE Eng Med Biol Soc, 2006. **1**: p. 37.
10. Butcher, E.C., E.L. Berg, and E.J. Kunkel, *Systems biology in drug discovery.* Nat Biotechnol, 2004. **22**(10): p. 1253-9.
11. Drews, J., *Stategic trends in the drug industry.* Drug Discov Today, 2003. **8**(9): p. 411-20.
12. Cohen, M.H., et al., *United States Food and Drug Administration Drug Approval Summary: Gefitinib (ZD1839; Iressa) Tablets.* Clin Cancer Res, 2004. **10**(4): p. 1212-1218.
13. Liebler, D.C. and F.P. Guengerich, *Elucidating mechanisms of drug-induced toxicity.* Nat Rev Drug Discov, 2005. **4**(5): p. 410-20.
14. Apic, G., et al., *Illuminating drug discovery with biological pathways.* FEBS Lett, 2005. **579**(8): p. 1872-7.
15. Kuhn, M., et al., *Large-scale prediction of drug-target relationships.* FEBS Lett, 2008.
16. Strong, M. and D. Eisenberg, *The protein network as a tool for finding novel drug targets.* Prog Drug Res, 2007. **64**: p. 191, 193-215.
17. He, X. and J. Zhang, *Why Do Hubs Tend to Be Essential in Protein Networks?* PLoS Genetics, 2006. **2**(6): p. e88.
18. Yu, H., et al., *The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics.* PLoS Comput Biol, 2007. **3**(4): p. e59.
19. Albert, R., H. Jeong, and A.L. Barabasi, *Error and attack tolerance of complex networks.* Nature, 2000. **406**(6794): p. 378-82.
20. Ekman, D., et al., *What properties characterize the hub proteins of the protein-protein interaction network of Saccharomyces cerevisiae?* Genome Biology, 2006. **7**(6): p. R45.

21. Tew, K.L., X.L. Li, and S.H. Tan, *Functional centrality: detecting lethality of proteins in protein interaction networks.* Genome Inform, 2007. **19**: p. 166-77.
22. Goh, K.I., et al., *The human disease network.* Proc Natl Acad Sci U S A, 2007. **104**(21): p. 8685-90.
23. Wishart, D.S., et al., *DrugBank: a comprehensive resource for in silico drug discovery and exploration.* Nucleic Acids Res, 2006. **34**(Database issue): p. D668-72.
24. Gandhi, T.K., et al., *Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets.* Nat Genet, 2006. **38**(3): p. 285-93.
25. von Mering, C., et al., *STRING: known and predicted protein-protein associations, integrated and transferred across organisms.* Nucl. Acids Res., 2005. **33**(suppl_1): p. D433-437.
26. Mishra, G.R., et al., *Human protein reference database--2006 update.* Nucl. Acids Res., 2006. **34**(suppl_1): p. D411-414.
27. Joshi-Tope, G., et al., *Reactome: a knowledgebase of biological pathways.* Nucl. Acids Res., 2005. **33**(suppl_1): p. D428-432.
28. Lee, H.K., et al., *Coexpression Analysis of Human Genes Across Many Microarray Data Sets.* Genome Res., 2004. **14**(6): p. 1085-1094.

Ragini Pandey

Address: 3443 Burlingame Blvd, Westfield 46074, IN

Email: rapandey@iupui.edu , ragini.pandey@gmail.com

Phone: (317) 873-3435

Objective

Work with an organization to learn and grow in terms of experience and knowledge and at the same time contribute and play my part in helping the organization achieve its goals.

Computer Skills

Programming languages: C#, Excel macro, PL/SQL, Perl, Python, Java, JDBC, MATLAB, XML, Assembly Language.

Web programming: ASP.NET, familiarity with Ajax.

Databases: MySQL, Oracle, familiarity with SQL server.

Operating systems: Windows XP, Vista, familiarity with Linux

Bioinformatics Skills

- Analyzing Drug-Target relationship
- Analyzing drugs and there side effects
- Database Queries and Sequence Retrieval
- Homology Searching
- Phylogenetic Trees Reconstruction: UPGMA algorithm, Neighbor-joining algorithm, maximum likelihood
- Parsing data using Python and Perl
- Data analysis.

Product Development

- ***Emergency Data Link*** - *Currently in market:* ***Windows Application, which converts the health information entered by the users into images and uploads into photo viewer.***

  The program allows users to enter their health specific information and save it to their local computer. For user privacy, it encrypts the entered information. Information is

then converted into high quality JPEG images and uploaded to the photo viewer. It generates a PDF report of the user information for printing purpose.

Software skills used:

- **C#**

Role:

- Eliciting and analyzing requirements
- Collaborating with device manufacturer to finalize design specification for device interface
- Preparing Design specification
- Primary developer
- Collaborating with external professional testing company to test the program

Website Development-

Developed an internal web application for Capital Equipment and Machinery Carmel, where the users can log in to update the machine and the client information. Also was involved in updating and maintaining their software.

Software skills used:

- **ASP.Net**
- **C#**

Experience – 2 yrs

Research Assistant

IUPUI (Indiana University Purdue University, Indianapolis)

Jan 2007- present

Product development

National Safety Net, Inc. -June – February 2008

Capital Equipment and Machinery- August 2006 – May 2007

Thesis Work

A Novel Network Biology Approach To drug Target selections-

In this study we are trying to see if there exists some relationship between drug targets, the essentiality of proteins and the network hubs. Does the degree of connectivity plays some role in selecting drug targets or whether it should be a factor. Predicting 'Good targets' and 'Bad targets' based on the side effects of drugs and the connectivity of proteins. We were able to find the 'good and bad targets' based on network analysis.

Skills used:

- **C#**
- **Excel macros**
- **Oracle,PL/SQL**
- **Graphs excel**
- **Minitab**
- **Python**

Project Work

- ***Developed a search engine for querying gene-protein information for the related micro-array affemetrix ids-***
  The user can query by entering one of the following descriptors: micro-array id, Gene ID, Protein ID. The user can also select for the needed information. We integrated the gene data from Entrez GENE, protein data from UNIPROT and interaction data from DIP. We also provided the link to KEGG for related pathways.

- ***Comparing GNOMES via orthogonal methods*****:** By investigating the similarities between Gene associations predicted via Genomic vs. Annotation similarity.
  We decided to work on two species. The sequences were downloaded from the NCBI and the sequence similarity search was performed using BLAST. We used Perl to automate the process of running BLAST against the sequences. The annotation data was downloaded from NCBI and TIGR. Integration techniques were used in order to incorporate the data from two different sources. MYSQL was used to store all the data. Comparisons between the results obtained by BLAST and the annotations were done based on the GO ids. A gene specific application was created to give a view of the results of the similarities for two genes from the two different methods.

Software skills used:

- VB.NET
- Integration techniques for biological data
- Excel macro
- Oracle, PL/SQL
- Perl, Python for parsing the data.

Education

Masters in Bioinformatics:

Aug 2006 – present

Current GPA: 3.8/4

Indiana University, IUPUI IN

Computer Courses

GPA: 3.8/4

Aug 2004– Dec 2005

North Carolina State University, NC

M.A in English

May1997-May 1999

Kanpur University, India

Bachelors in Biology

May 1992-May1995

Allahabad University, India

May 1992-May1995

Computer Course Work

Java Programming

Data Structures

Discrete Mathematics

Assembly Language Programming

Object oriented analysis and design

Bioinformatics course work:

Introduction to Informatics

Introduction to Bioinformatics

Biostatistics

Structural Bioinformatics

Data Integration

Research Design

Mach learning and Pattern Recognition

References

- Dr. Jake Chen
  Department of Bioinformatics IUPUI

  Contact No: 317-278-7604

- Paul Kaufmann
  National Safety Net, Inc. Carmel.

  Contact No: (317) 587 0438

- Dr Pedro Romero
  Department of Bioinformatics IUPUI