INTERACTOME NETWORK ANALYSIS TO IDENTIFY DRUG TARGETS OF *MYCOBACTERIUM TUBERCULOSIS H37RV*

A THESIS

Submitted in partial fulfilment of the requirements for the award of the degree

of

DOCTOR OF PHILOSOPHY

in

MATHEMATICS

by

TILAHUN MELAK SITOTE

DEPARTMENT OF MATHEMATICS
INDIAN INSTITUTE OF TECHNOLOGY ROORKEE
ROORKEE-247667 (INDIA)
APRIL, 2016

INDIAN INSTITUTE OF TECHNOLOGY ROORKEE
ROORKEE

CANDIDATE'S DECLARATION

I hereby certify that the work which is being presented in the thesis entitled "**INTERACTOME NETWORK ANALYSIS TO IDENTIFY DRUG TARGETS OF *MYCOBACTERIUM TUBERCULOSIS H37RV*"** in partial fulffilment of the requirements for the award of the Degree of Doctor of Philosophy and submitted in the Department of Mathematics of the Indian Institute of Technology Roorkee, Roorkee is an authentic record of my own work carried out during a period from August, 2012 to April, 2016 under the supervision of Dr. Sunita Gakkhar, Professor, Department of Mathematics, Indian Institute of Technology Roorkee, Roorkee.

The matter presented in this thesis has not been submitted by me for the award of any other degree of this or any other Institute.

(Tilahun Melak Sitote)

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

Sunita Gakkhar
(Supervisor)

The Ph. D. Viva-Voce Examination of **Tilahun Melak**, Research Scholar, has been held on **April 26, 2016**.

Chairman, SRC

Signature of External Examiner

This is to certify that the student has made all the corrections in the thesis.

Signature of Supervisor
Dated: 29/04/2016

Head of the Department 29/4/2016

Abstract

The preliminary step in rational-based drug discovery is identifying the society's unmet medical needs that are not properly addressed with the available treatments. After prioritizing the unmet medical needs, drug target identification is the first and key phase in the pipeline. In the previous target-based drug discovery the failure rate is very high. Many of these failures are attributed to improper target identification. Our inadequate knowledge about the disease and molecular mechanisms played a significant role. The wealth of data and information in this 'omics' era present immense of new opportunities to enhance our understanding about the disease dynamics at cellular and molecular level. With these advancements, the task of successful identification of therapeutic targets becomes more promising. However, there is no single sufficient-enough method due to the complexity of human diseases, heterogeneity of biological data and inherent limitations of each approach. Therefore, systematically integrated computational methods can be used to identify potential drug targets for high burden drug-resistance diseases like tuberculosis (TB).

TB is an infectious disease caused by the infamous etiological agent *Mycobacterium tuberculosis* (Mtb). It is the cause of morbidity and mortality to millions every year. *Mycobacterium tuberculosis H37Rv* is the most studied strain of TB. The emergences and rise of drug-resistance is the main bottleneck for the management, control and eradication programs of TB. Various strategies have been implemented to counter the problem of resistance. However, available statistics indicates that resistance forms are still on the rise. Drugs used in the current treatment of drug-resistance TB are expensive, toxic, with adverse side effects and ineffective to act on the latent forms of bacillus. The stated shortcomings highlight the requirement of new therapeutic targets.

In this thesis, comprehensive protein-protein interactome network analyses have been carried out to identify potential drug targets and co-targets of *Mycobacterium tuberculosis H37Rv.* Proteins involved in the same cellular processes often interact with each other. Protein–protein interaction network analysis is fundamental to understand the complexity of biological systems by revealing hidden relationships between drugs, genes, proteins and diseases. There is enormous amount of protein-protein interaction data in various repositories due to the advancement of techniques such as two-hybrid systems, mass spectrometry, and protein microarrays. These analyses have been carried out by aiming to obtain important system-level insights about TB and counter the challenges at the target identification phase of drug discovery process. *In silco* molecular

modelling and structure analysis has also been carried out for protein translocase subunit SecY (Rv0732).

The list of potential primary drug targets has been identified through analysis of comparative genome and network centrality measures on the protein-protein interaction network of the pathogen. The interaction dataset was retrieved from STRING. It is one of the main sources of protein-protein interaction data of TB. It acts as a meta-database by integrating interactions from numerous sources such as experimental repositories, computational prediction methods and public text collections. The protein-protein interaction dataset of *Mycobacterium tuberculosis H37Rv* in STRING has been shown that it is of low quality by containing false positives and false negatives. This can affect the results of any analysis which is based on this dataset. To minimize the impact, the portion of the dataset which is more reliable has been considered. The four centrality measures degree, closeness, betweenness and eigenvector have been used to identify the most central proteins in the interactome network. Only proteins that found at the centre of gravity of the interactome network were considered. BLAST search of protein coding genes has been carried out against DEG to filter out genes which are essential for the survival and growth of the pathogen. The corresponding protein sequences obtained after DEG search were subjected to BLASTp search against the non-redundant database with an e-value threshold cut off set to 0.005 and restricted to Homo sapiens to avoid the possible host toxicity at the sequence level. A list consisting of 137 proteins have been proposed as potential primary drug targets of *Mycobacterium tuberculosis H37Rv*. These proteins are believed to be reliable targets since they are reported as essential proteins for the growth and survival of the pathogen, have no detectable homology with human so as to prevent host toxicity and prioritized based on their network centrality measure values where all of them were found within the close neighbourhood of the centre of gravity of protein-protein interaction network. Many of the proteins in the list have been reported as drug targets by other methods.

The potential primary drug targets have been further prioritized based on their influence to resistance genes using maximum flow approach on weighted proteome interaction network of the pathogen. The weighted protein-protein interaction network of the pathogen has been constructed using a dataset retrieved from STRING. The combined score values of the pair of interacting proteins has been assigned as weight of interactions. The potential drug targets and resistance genes have been taken as inputs. Then, the potential drug targets have been prioritized based on their maximum flow value to resistance genes. This approach does not suffer from biasness towards shortest paths since it is based on flow. More importantly, the inhibition of a

protein which has more influence on the resistance genes of the existing drugs is expected to disrupt the communication to these genes. Hence, it can be considered as an additional druggablity assessment criteria for drug resistance diseases like TB.

Our limited system-level knowledge about the possible routes of resistance is one of the causes of failure to strategies against drug-resistance TB. Detailed analysis has been carried out to explore these routes through which information required for triggering drug-resistance may be passed on in the cell. Proteins involved in the emergence of resistance by mediating information among drug target proteins of eight clinically used drugs in the current treatment regime of TB and resistance genes have been identified. These lists of proteins have been proposed as potential co-targets of each drug. The analysis has been carried out on weighted drug-specific protein-protein interaction networks of the pathogen. The validated drug targets and resistance genes have been taken as inputs. The maximum flow values of proteins in the flow from validated drug targets to resistance genes have been computed. Proteins have been prioritized based on their maximum flow value. Subsequent filters such as non-homologous assessment to avoid host toxicity, identification of proteins that interact with the host and essentiality analysis have been carried out. The final refined lists of proteins have strong involvement in the emergence of drug resistance and targeting them with systematic combination of existing drugs is believed to be effective to prevent the emergence of drug resistance.

In silco structural analysis of protein translocase subunit SecY (Rv0732) has been carried out to get descriptive three-dimensional structure. Rv0732 has been selected because it is highly ranked potential drug target without solved three-dimensional structure. The active site has been identified for protein-ligand or protein-inhibitor binding.

Acknowledgements

First and foremost I am very thankful to almighty God who has blessed me with a lot of things. It seemed a very tough journey and it actually is but with the sake of God it has turned out to be great. It is a great pleasure for me to be formally very thankful to all the people who are around me to achieve what I have achieved. I am very lucky to be surrounded by great individuals both in family and in professional life. Their contributions in shaping my life and this wok have been immense. I might have walked the hard walk but I was not alone in doing so.

My Guide, Professor Sunita Gakkhar, was very instrumental for the success of this work. I have learnt a lot from her rich experiences of teaching, research and supervising. There were a lot of instances where I have been amazed about how she made things look easy. I consider myself lucky to have such a great advisor who understands me very well and constantly encourages me to put a lot of hard work throughout the course of my stay in the Institute. This work has been materialised due to her insightful directions.

I sincerely acknowledge Professor V.K Katiyar, HOD of Mathematics, Professor Kusum Deep, SRC chair and all the SRC members for facilitating the required procedures very smoothly during the course of this thesis.

For me to be who am, the biggest credit goes to my parents; my father, Melak Sitote, and my mother, Workye Semagn. There are a lot of things that I have learnt from them and definitely I will also continue to learn. The most important is that they showed me how to be disciplined and humble. They showed me that no matter how tough the situations are there is always a way out as far as we use our mind. They might not have all the resources but they were trying to be sure that I would have what I needed. Their wisdom is always their main tool to guide me. I am nowhere near to pay them back for what they have done to me but I hope I made them proud with what I have achieved. I will never forget the inspirations I have got from my aunt, Biritu Semagn, during the earlier stages of my education. In fact, I hate to call her my aunt because she is like a mother to me. She has a better understanding about the value of education than anyone I know.

My beloved wife, Hiwot Taddesse, was the main part in this journey. Being in distance from your husband and raise a kid by yourself is high up there in terms of the toughest things you can possibly face in life. But she handled it remarkably well. She is the source of inspiration for me when things do not go in my way. I have learnt a lot about how to be committed for my objectives and long term life goals. I am very lucky to have her by my side and I cannot wait to be with her to share all the happiness and hardships of life.

My dear sister, Serkalem Haile, is the first person who comes to my mind whenever I have faced problems or got some good news. She is the most dependable, generous and problem solver in the family. She is not only my sister she is also my best friend. Family is the most important thing and I am blessed with a great one. I am very thankful for all of them from the bottom of my heart.

I am also very thankful for all of my friends and acquaintances. With friends, I have shared a lot of happiness and got experiences that will be very valuable to my future. Even from the people who do not have good attitude about me and have some disagreements with, I have learnt a lot about patience and be able to build character. They thought me that everything is not black and white where there are times that I have to stand up for myself.

Finally, I would like to acknowledge the financial support by Ministry of Education of Ethiopia (MoE) and all the required facilities I got from the Department of Mathematics, IIT Roorkee.

(TILAHUN MELAK)

Roorkee
April, 2016

List of Publications

Refereed Journals

1. Melak T, Gakkhar S. **Potential non homologous protein targets of mycobacterium tuberculosis H37Rv identified from protein–protein interaction network.** *J. Theor Biol. 2014*, 361: 152–158. doi: 10.1016/j.jtbi.2014.07.031.(Elsevier)
2. Melak T, Gakkhar S. **Maximum flow approach to prioritize potential drug targets of Mycobacterium tuberculosis H37Rv from protein-protein interaction network.** Clin Transl Med. 2015; 4(1):61. doi: 10.1186/s40169-015-0061-6. Epub 2015 Jun 5. (BioMed Central, Springer)
3. Melak T, Gakkhar S. **Comparative Genome and Network Centrality Analysis to Identify Drug Targets of *Mycobacterium Tuberculosis H37Rv*.** Biomed Res Int. 2015; 2015: 212061. doi:10.1155/2015/212061(Hindawi)
4. Melak T, Gakkhar S. **Identifying potential co-targets of *Mycobacterium tuberculosis H37Rv* using maximum flow approach.** (Communicated).
5. Melak T, Gakkhar S. **Structural analysis of protein translocase subunit SecY from *Mycobacterium tuberculosis H37Rv*: a potential target for anti-tuberculosis drug discovery** (communicated).

Conference

1. Melak T, Gakkhar S. **Potential Co-Targets of Isoniazid Identified through Proteome Interactome Network Analysis,** presented at International Conference on Current Challenges in Drug Discovery Research (CCDDR 2015), Malviya National Institute of Technology, Jaipur-302017, Rajasthan, India. November 23-25, 2015.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

List of Abbreviations

AD	Alzheimer Disease
AMK	Amikacin
BLAST	Basic Local Alignment Search Tool
CAP	Capreomycin
CIP	Ciprofloxacin
CS	Cycloserine
DEG	Database of Essential Genes
EREG	Epiregulin
ESPript	Easy Sequencing in PostScript
ETH	Ethionamide
FIFO	First In First Out
FLD	First-line drug
GO	Gene Ontology
HGT	Horizontal Gene Transfer
HIV	Human Immunodeficiency Virus
IC	Information Centrality
INH	Isoniazid
KAN	Kanamycin
LAC	Local Average Connectivity-based method
LEV	Levofloxacin
MCBS	Minimum Common Bioactive Substructure
MDR-TB	Multidrug-resistant Tuberculosis
MOA	Mode of Action
MOX	Moxifloxacin
Mtb	Mycobacterium Tuberculosis
MTBC	Mycobacterium Tuberculosis Complex
NCBI	National Center for Biotechnology Information
NC	Network Centrality
NADH	Nicotinamide Adenine Dinucleotide
NEI	Neuro-Endocrine- Immune
OFX	Fluoroquinolones Consists of Ofloxacin
OSDD	Open Source Drug Discovery

PAS	P-aminosalicylic Acid
PDB	Protein Data Bank
PTH	Prothionamide
SAR	Structure-Activity Relationships
SLD	Second-Line Drug
STRING	Search Tool for the Retrieval of Interacting Genes/Proteins
SVM	Support Vector Machine
SC	Sub-graph Centrality
TB	Tuberculosis
TDR	Tropical Disease Research
TNF	Tumor Necrosis Factor
TraSH	Transposon Site Hybridization
XDR-TB	Extensively Drug-Resistant Tuberculosis
VEGF	Vascular Endothelial Growth Factor
WHO	World Health Organization
Y2H	Yeast Two-Hybrid

CHAPTER ONE

INTRODUCTION

1.1. Drug Discovery

Drug is a chemical substance that can be used in the cure and prevention of endogenous diseases. The main causes of endogenous diseases are infectious vectors such as virus, bacterium and parasite. In some cases, in-born sequence errors in germ cells or spontaneous mutations in somatic cells can be the possible causes of these diseases. Drug discovery is a process consisting of a number of phases through which new potential medications are being identified. The path of drug discovery starts from identification of unmet medical need. Usually, there are various unmet medical needs in a society that require medical attention such as cases with no or limited treatment options and ineffectiveness of available medications due to the emergence of drug resistance. One of the main objectives of drug discovery is to fulfil those unmet medical needs. Following the effective identification of unfulfilled health needs, druggable biological targets that could relieve the symptoms of the disease, or, as in the recent years, that are involved in the causative process of the disease will be identified [1-3].

In the past, drug discovery was more of serendipitous discovery where active ingredients are identified from traditional remedies followed by screening for chemical libraries of synthetic small molecules, natural products or extracts in intact cells or whole organisms to identify substances that have a desirable therapeutic effect in a process known as classical pharmacology. The use of high throughput screening has become a common practice since human genome were sequenced which makes rapid cloning and synthesis of large quantities of purified proteins become a reality. In reverse pharmacology, large compounds libraries screened against biological targets that are believed to be disease modifying in the process. As interdisciplinary endeavour with an industrial base, drug research is not much older than a century [4].

The drug discovery process is very expensive, lengthy and complicated. For instance, the development of a new prescription medicine that gains marketing approval could need a total cost of more than 700 Million dollars [5]. From the earliest stages of discovery to the time it is available for treating patients could also take more than a decade. More importantly, after such huge investments of time and money, the likely hood of the resulted drug to be failed is significantly high. Incredible wealth of knowledge has been generated over the last two decades due to tremendous growth in computational power, advancements in genomics and proteomics along with the explosion of biological data. Due to these advancements, it is possible to understand the disease dynamics at cellular and molecular level so that the task of discovering and developing safe and effective drugs becomes more promising. This minimizes not only the risk of failures in drug discovery process but saves time, effort and money.

1.1.1 Drug Discovery Pipeline

The process of drug discovery can broadly follow three different paradigms [6]:

- Physiology-based drug discovery
- Target-based drug discovery
- Function-based drug discovery

Physiology-based approach is a traditional drug discovery paradigm in which the organism is seen as a black box and drugs are characterized on the basis of their physiological effects in complex disease-relevant animal models [7]. It had been successful in the discovery of many effective treatments. The approach is still in use intensively for the development of antipsychotic and antidepressant drugs. Physiological approach doesn't require the knowledge of aetiology and mechanism of action of the disease and because of that, it has a difficulty of clearly defining the relationship between the drug mechanism of action and biological effects. It also suffers from a very low-throughput screening capacity.

Physiological-based approach was replaced by target-based drug discovery paradigm in the last two decades. Target-based approach is a more rational drug discovery paradigm that allows increased screening capacity. The road of target discovery starts from identification of therapeutic targets and their role in the disease unlike physiological approaches where target identification/validation step are being forgone to start from screening. Target identification could be a very difficult task due to the existence of thousands of human or pathogen genes and the variety of their respective gene products. Additionally, insight into the "normal" or "native" function of a gene or gene product does not necessarily connect the gene or gene product to disease. There was high expectation of success from the target-based drug discovery approach but the final outcome is debatable [8, 9].

In functional approach the objective is to persuade therapeutic effect through applying mechanisms of action to disease-specific functional abnormality [7]. Compounds are screened based on their ability to induce or normalize functional parameters in disease-relevant models like axonal transport, growth processes, hormone secretion or apoptotic processes. Since function requires the integrated action of many mechanisms, functional parameters represent a higher level of organism complexity when it is compared with target-based approach. Micro-dialysis and whole-cell or extracellular electrophysiology are practical examples of function-based drug discovery approach. The main limitation of this approach is its screening capacity where it is very low and cannot be used for library screening.

Target-based drug discovery approach has been the dominated paradigm since the early 1990's mainly due to its implementation strategy where it follows a scientific approach by defining specific molecular mechanism or mode of action [7]. In theory, it seems an approach which can lead to successful discoveries of noble medications. However, there is significantly huge deviation between theoretical assumptions and results where the identified targets have only 3% of success rate of reaching to preclinical development [10]. This couldn't be acceptable for the pharmaceutical companies who invest several hundred million dollars and spend decades for the production of a single drug. So, it is understandably the reason which makes them less interested in potential drug targets identified through various investigations. There are probably a number of reasons and explanations for this failure but the main reason behind it is our inadequate level of understanding about the disease and its biological process to predict the therapeutic value or druggability of a novel target. The pipeline followed by target-based drug discovery approach is very similar across different pharmaceutical companies [3]. It comprises the following five detail steps which have also been shown in Figure 1.1:

1. *Target identification and validation:* - druggable biological targets are identified from biological and clinical findings. Then, they will be validated through expression patterns and knockout mice if a selective compound is not available. In most cases drugs are designed to inhibit a target under consideration and usually they are effective. But there should be an absolute certainty about the essentiality of the target in a given disease ahead of proceeding to the next step of drug discovery pipe line. The complete verification can be obtained by testing the idea in human which is not possible especially at the initial stages of drug development. Thus, the target has to undergo rigorous validation to clearly define its role in specified disease.
2. *Hit identification:* - validated targets are screened against a library of ten to hundreds of thousands of compounds through high throughput screening to identify those compounds that hit targets [11]. The numbers of compounds identified from this step are usually hundreds in number and they will be filter out in the *lead identification* step [3].

3. *Lead identification*: - In order to filter out the identified lead compounds, novel compounds which contain active substructures are synthesized by analyzing the structure of the selected compounds and identifying common active substructures [3].
4. *Lead optimization*: - the identified leads are further screened for target selectivity and pharmacokinetic properties such as absorption and bioavailability [3, 6]. The screening is also aimed to increase their potency and efficacy, while decreasing side effects and toxicity. Knowledge of the mode of action (MOA) can be useful in simplifying *lead optimization* process through predicting the effect of drug interactions, thus allowing structure-activity relationships (SAR) to guide medicinal chemistry efforts toward optimization [11]. As helpful as it could be, the limitation with MOA is that targets of many drugs are unknown and it is difficult to find them from thousands of gene products [3].
5. *Development:* - Compounds with positive results in an *in vivo* disease model for proof of principle from the screening process of lead optimization step will be selected for development.

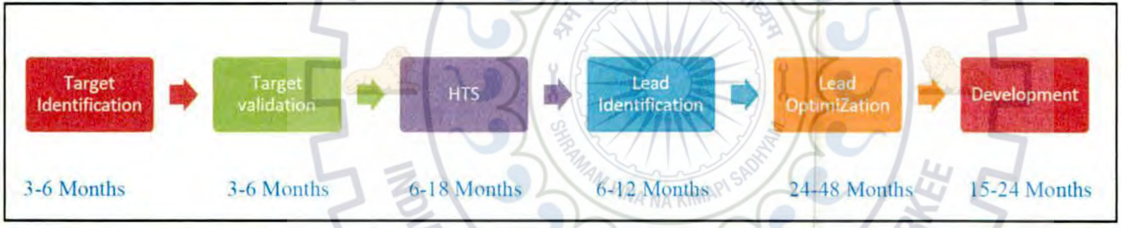


Figure 1.1 Drug discovery pipeline

1.1.2 Computational Techniques in Drug Discovery

Drug discovery is a multi-disciplinary endeavour which involves a wide range of scientific disciplines including biology, chemistry, pharmacology, statistics and bioinformatics. Computational biology and Bioinformatics can have a significant impact in current regime of drug discovery process by substantially speeding up the process to reduce the costs and even changing the way drugs are designed [2]. Through time, their involvement and importance has significantly increased spreading across all aspects of drug discovery, drug assessment, and drug development [12]. As it can be expected there is a considerable growing in importance of bioinformatics tools in handling large volumes of data due to explosions of biological data in this 'omic era'. Their role in predicting, analyzing and interpreting clinical and preclinical findings has also shown a significant improvement.

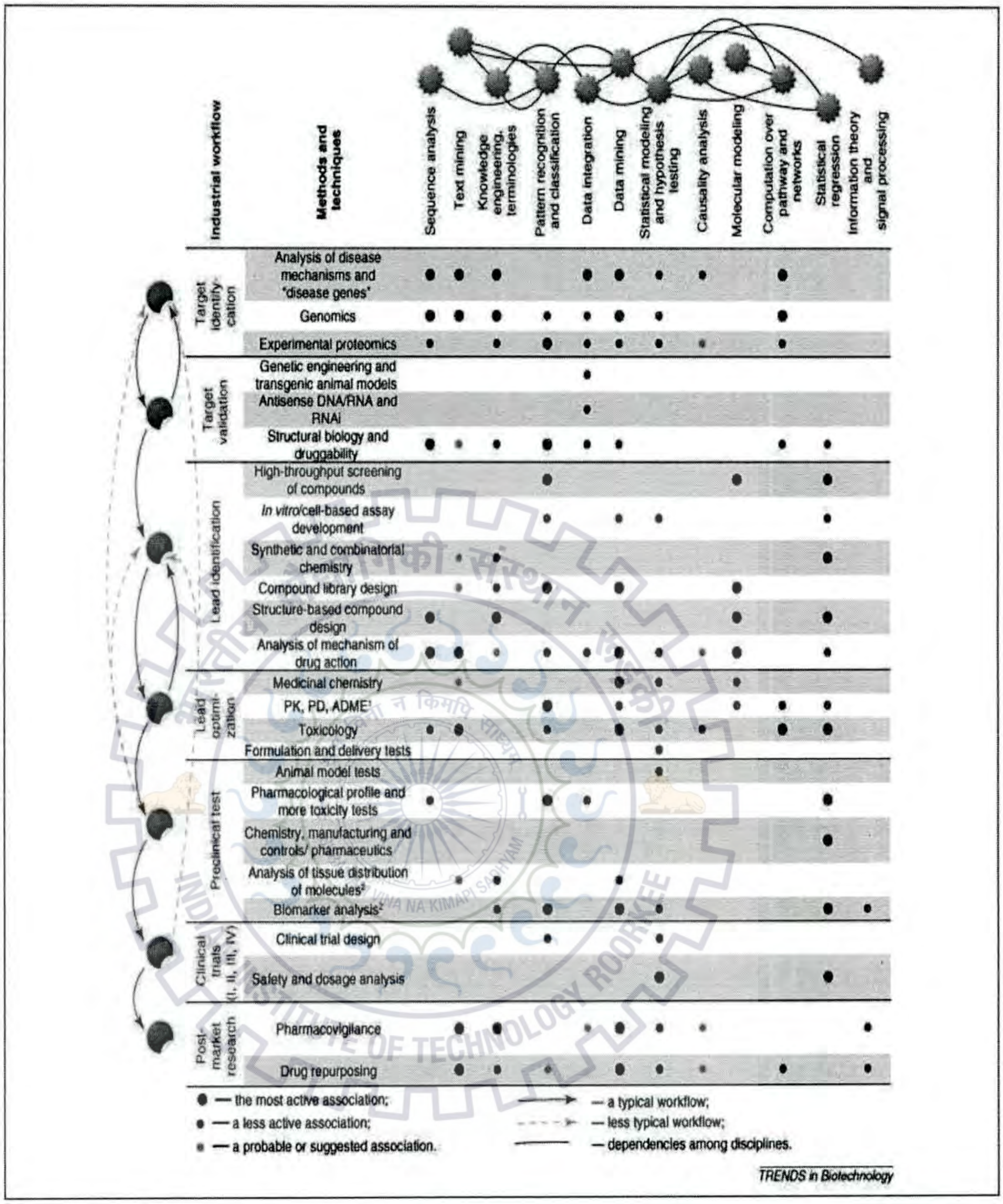

Industrial workflow
Methods and techniques
Sequence analysis
Text mining
Knowledge engineering, terminologies
Pattern recognition and classification
Data integration
Data mining
Statistical modeling and hypothesis testing
Causality analysis
Molecular modeling
Computation over pathway and networks
Statistical regression
Information theory and signal processing
Target identify-cation
Analysis of disease mechanisms and 'disease genes'
Genomics
Experimental proteomics
Target validation
Genetic engineering and transgenic animal models
Antisense DNA/RNA and RNAi
Structural biology and druggability
Lead identification
High-throughput screening of compounds
In vitro/cell-based assay development
Synthetic and combinatorial chemistry
Compound library design
Structure-based compound design
Analysis of mechanism of drug action
Lead optimi-zation
Medicinal chemistry
PK, PD, ADME1
Toxicology
Formulation and delivery tests
Preclinical test
Animal model tests
Pharmacological profile and more toxicity tests
Chemistry, manufacturing and controls/ pharmaceutics
Analysis of tissue distribution of molecules2
Biomarker analysis2
Clinical trials (I, II, III, IV)
Clinical trial design
Safety and dosage analysis
Post-market research
Pharmacovigilance
Drug repurposing
● — the most active association;
• — a less active association;
• — a probable or suggested association.
— a typical workflow;
— less typical workflow;
— dependencies among disciplines.
TRENDS in Biotechnology

Figure 1.2 Role of computational technologies in the drug discovery pipeline [12]

The figure is taken from Yao et al. (2009) and it shows a comprehensive description about the involvements of various methods of computational biology in different stages of drug discovery process. It has been indicated that the traditional linear steps are increasingly being shifted to more of iterative and parallel steps in pharmaceutical companies to increase productivity.

The roles of computational biology in each steps of drug discovery process were effectively summarized in the comprehensive review by Yao et al. (2009) [12]. As shown in Figure 1.2, different computational methods are commonly used throughout the drug discovery pipeline for mining knowledge from different types of biological data, building network models of molecular processes, statistical and causal analyses. These methods include Sequence analysis, molecular modeling, simulation of molecular networks, probabilistic data integration, and development of drug cocktails. Sequence analysis is an indispensable part of target identification phase that is used to compare nucleotide and amino acid sequences. This will allow researchers to impute gene's function by considering evidence from homologous genes. The emerging computational approaches are promising in broadening our understanding of biological systems particularly about diseases; thereby novel therapeutic strategies could be designed.

Text mining is one of the emerging disciplines attracting a lot of interest in the biomedical sciences in recent years mainly due to the availability of an overwhelming amount of biomedical knowledge recorded in texts and its ability to identify, extract, manage, integrate, and exploit this knowledge [13-16]. It can discover new, hidden, or unsuspected knowledge from unstructured data. Currently, screening articles for biological terms including molecule names and biological statements such as molecular interactions is the common application area of text mining. It is very useful for efficient and systematic collection, maintenance, interpretation, curation, and discovery of knowledge.

1.1.3 Challenges in Drug Discovery

Drug discovery is a very expensive industry that requires pharmaceutical companies to spend several hundred millions to billions of dollars for the production of a single drug. It is also very slow where the entire process could take up to a decade. The worst part is that such huge investment of time and money could be fruitless due to a very low success rate of the so called noble targets to go all the way up to approval [10]. In fact, it has been stated that the estimated success rate of a selected target to reach even to a pre-clinical stage of the drug discovery pipeline is 3%. This is even without considering attrition rates caused by toxicological effects and lack of clinical effects. The very low success rate of new potential targets has substantially decreased the interest of pharmaceutical industries and instead they have changed their focus towards known targets and existing drugs which have an estimated success rate of up to 17% [6, 10]. However, due to the existence of limited possibilities, such option could only be a temporary solution.

There are various sources of error in the discovery of drugs from new potential targets through target-based drug discovery paradigm [12]. The main reason is our limited level of insight about the disease and molecular mechanisms underlying complex human phenotypes [17, 18]. Our understanding could be greatly enhanced with the abundantly available biomedical data and information. However, these data and information are available in heterogeneous formats spreading across various disintegrated repositories [19]. The integration of information from these sources is a major challenge that could require significant methodological and even cultural changes in our approach to data. There is also lack of suitable models which are capable of accurately simulating the link among successful discovery, organization of researchers and resources that underpins it.

The emergence of various forms of resistance to the existing drugs and adverse drug reactions are among the main emerging bottlenecks for drug discovery industry [20]. In spite of spending tremendous amount of money and time, there is an increase in withdrawal of high-profile drugs with fewer approvals of new drugs. Drug discovery as a multi-disciplinary endeavor requires integration of advanced scientific knowledge and resources from various disciplines. However, the previous drug discovery approaches are closed-door and market driven where there exist very limited collaborations. The overall implication of all of the stated challenges and shortcomings directed towards the urgent requirements of change in the current drug discovery system that is very expensive, has low efficacy, and with high adverse drug reactions. Open Source Drug Discovery (OSDD) project, initiated by Council of Scientific and Industrial Research, India is a promising paradigm to integrate global efforts of drug discovery [21]. There is an ideal suggestion to develop personalized medicine that treats whole systems and brings the right drug to the right patient with the right dosages [20]. Of course it is very easy to say than implement.

1.1.4 Target Discovery

A preliminary comprehensive analysis is required as a pre-drug discovery step to have detail understanding about the disease to be treated. This can be helpful in increasing our level of understanding about the diseases and its process from various perspectives such as its underlying causes of the condition, the mechanism of gene alteration and their effect on the proteins encoded by those genes, the way of protein-protein interactions among those proteins, the mechanism of how those affected cells change the specific tissue they are in and the overall effect of the disease on the patient. After being able to understand the disease and its process, a potential drug target will be identified. The term drug target in biomedical science represents a very broad concept which could be a molecular entity including gene, protein, miRNA, or it could be biological phenomena like molecular functions, pathways and phenotypes as far as it is relevant to a specific disease and its progression [22]. A target of interest could also be either a human molecule or a part of pathogen's cellular machinery. In the former case, a drug is designed to recognize and modify the target in order to achieve an intended therapeutic effect and in the latter case, its role is to kill the pathogen by interrupting the target.

The current drug discovery process requires following very sophisticated steps, consumes tremendous amount of time and money with a very high failure rate. The main causes of failure for previous drug development are largely attributed to improper target selection [7, 22, 23]. This makes target identification a key step in the drug discovery pipeline. Moreover, being able to reduce failure at the initial stages of the process is more effective than filling a pipeline with poorly chosen late-stage products that are very likely to fail, and fail expensively [24].

The two approaches for target discovery process are: system and molecular approach [22]. The target in a system approach is identified from the study of a disease of interest through clinical trials and *in vivo* animal studies. In molecular approach, potential druggable targets discovered based on information derived from interactions with small molecules. Molecular approach is the common paradigm that has been widely used in the current target discovery process [23, 25]. System level understanding of disease phenotypes through cellular interaction network is also important in addition to identifying, prioritizing and selecting reliable targets for the sake of identifying predictive models and constructing biological networks for human diseases [22]. An extensive collection and organization of multitude of heterogeneous data and information will be helpful to have a comprehensive understanding.

In this 'omics' era, there is an explosion of biomedical data and information [26]. For instance, more than 18 million abstracts were stored in MEDLINE/PubMed alone where 60,000 new abstracts are added monthly. There is an enormous growth in repositories of chemical, genomic, proteomic and metabolic data with an estimation of 100% increase every two years. This wealth of biological data and information presents immense new opportunities that have significantly boosted target identification in the target based drug discovery process to diagnose and fight human diseases [23]. With these immense of new opportunities, successful identification of therapeutic targets using computational approaches seems much feasible, requiring future experimental validation [24, 27]. However, no single method is sufficient enough due to complexity of human diseases, the heterogeneity of various biological data and the inherent limitations of various specific computational approaches [22]. So, it is timely to use integrated computational methods for an effective identification of potential drug targets. This approach could be applied on high burden drug resistance diseases like tuberculosis. In this thesis, systematically integrated computational techniques have been used to identify and prioritize potential drug targets and co-targets of mycobacterium tuberculosis H37Rv.

1.2. Literature Review

Tuberculosis (TB) is one of the main global public health threats of human where more than one-third of the world population is infected with it. It is caused by the infamous etiological agent *Mycobacterium tuberculosis* (*Mtb*) [28]. It particularly affects lung in the case of pulmonary TB and other organs of human body if it is extra pulmonary TB [29]. TB is communicable disease that spread through air when pulmonary TB infected people expel the bacteria with coughing. There are different clinical strains of TB and *Mycobacterium tuberculosis H37Rv* is the most studied strain [30]. The probability for the development of TB disease in a person infected with *Mtb* is usually small unless in cases like the specified person is also infected with human immuno-deficiency virus (HIV). TB is the second leading cause of death from an infectious disease, next to HIV [31]. According to the estimates of WHO global tuberculosis report of 2014, there were 9 million people who developed TB and 1.5 million who died from the disease in 2013 [29]. Out of those 1.5 million who died due to TB, 0.4 million were HIV-positive which shows that TB gets worse and more complicated on those who are infected with HIV. It is very difficult to accept any death from TB because most of the cases are curable if the infected people can access adequate health care diagnosis and correct treatment adherence.

The mechanism of worldwide TB diagnosis is more or less similar where bacteria are observed in sputum samples using sputum smear microscopy which was developed before 100 years [29]. Recently, the use of rapid molecular tests and culture methods in countries with more developed laboratory capacity to diagnose both TB and drug-resistant TB are increasing. Popular first-line drugs that can cure about 90% of new TB cases have been around for many decades. The first successful drug for the treatment of TB was developed in 1940. Rifampicin is the most effective first-line anti-TB drug till this day. It was introduced in the 1960s. The current treatment of new cases of drug-susceptible TB includes four first-line drugs: isoniazid, rifampicin, ethambutol and pyrazinamide with a six-month regimen.

One of the main reasons for the ineffectiveness of treatments with first-line drugs of TB is the emergence of multi-drug resistance tuberculosis (MDR-TB) and extensive drug-resistance tuberculosis (XDR-TB) [32, 33]. MDR-TB is resistance to at least isoniazid and rifampicin which are the two most powerful anti-TB drugs [34, 35]. XDR-TB, in addition to being resistance to at least isoniazid and rifampicin, is resistance to at least three of the six main classes of second-line drugs (SLDs) [36]. The WHO estimate showed that 3.6% of the new and 20.2% of previously treated TB cases have had MDR-TB and an estimated of 9.0% of patients with MDR-TB had XDR-TB in 2013 [29]. The resistance of TB is even getting worse with the emergence of Total drug-resistance tuberculosis (TDR-TB) in three countries; India, Iran, and Italy as it has been documented in four major publications [37-40]. Treatments of drug-resistance TB takes longer duration that can extend up to 20 months. The current second-line drugs for the treatment of drug-resistance TB are broadly categorised into two types called Fluoroquinolones and Injectable anti-tuberculosis drugs. Fluoroquinolones consists of Ofloxacin (OFX), levofloxacin (LEV), moxifloxacin (MOX) and ciprofloxacin (CIP). Kanamycin (KAN), amikacin (AMK) and capreomycin (CAP) are injectable drugs of drug-resistance TB. There are also few less effective second-line anti-tuberculosis drugs. These are Ethionamide (ETH)/Prothionamide (PTH), Cycloserine (CS)/Terizidone, P-aminosalicylic acid (PAS).

The existing front-line anti-mycobacterial drugs are mainly responsible for controlling the disease to the extent that exists today [41]. However, they have many shortcomings. As it has been stated, the main is the emergence of different levels of drug-resistance which could be able to render these front-line drugs inactive. Drugs that are used to treat drug-resistance TB are more expensive and more toxic than first-line drugs. Some of the drugs like rifampicin have adverse side effects which lead to patient compliance. Most of these drugs are also not effective to act on the latent forms of bacillus. The need for careful consideration of vicious interactions between TB and HIV during drug discovery process for *Mtb* extends the challenge further [42].

In order to encounter the problem of drug resistance, a number of different strategies are being implemented such as rotation of antibiotic combinations, enhanced medical supervision to ensure patient compliance, identification of new targets that may be less mutable, search for new chemical entities for known targets, use of virulence factors as targets and 'phenotypic conversion', which aims to inhibit the resistance mechanism employed by the bacterium [43]. Despite the trial of these important resistance measures and a lot of researches going on as well to understand the pathogenesis of Mycobacterium tuberculosis, available statistics indicates that resistance forms are still on the rise [44, 45]. This is strong indicator for the requirement of identification of new therapeutics for TB. Identification of potential new drug targets that can provide detail know-how of the infectious organisms for the development of effective drugs is the current interest of the scientific community involved in the discipline of drug discovery [27]. According to a recent WHO report, there are new and repurposed anti-TB drugs on the verge of successfully emerging from the tunnel of drug discovery pipeline and combination of new compounds are being tested in clinical trials [29]. Two new drugs, bedaquiline and delamanid, recently passed the approval and ready to be used for the treatment of MDR-TB under specific conditions. It has been also reported that there are many vaccines in advanced stages of clinical trials. This seems very good news after such a tiresome effort and long wait. Still, more researches and integrated efforts are required to attain various goals set to control TB.

The readily available enormous biomedical data and information led to exhaustive computational and experimental investigations for identification of new drug targets. One of the computational methods used in the identification of drug targets is commonly called random walk and it is based on traditional computational approach [46]. It utilizes structural information to predict whether a protein can be drug target or not [47]. Even though this method achieved reasonable performance, it suffers from limited availability of protein 3D structures. Recently, there are significant computational advancements in predicting structures with the aid of protein remote homology detection [48, 49]. The two mainly used approaches in the past two decades are generative methods and discriminative algorithms [48]. Since generative methods used only positive training samples to build the models for prediction, they were not effective. Unlike generative approach, the discriminative methods use both positive and negative samples in the process of training. Support Vector Machine (SVM) is most widely used, effective and accurate discriminative method for remote homology detection problem [50, 51]. The feature vectors incorporating the positional information of amino acids or other protein building blocks in Support Vector Machine based method was stated to be promising than position independent methods [48]. It is important to note that the profile-based approach specifically holds high potential in remote homology detection. In this regard, improved performance has been achieved by Combining Chou's Pseudo Amino Acid Composition and Profile-Based Protein Representations extracted from the frequency profiles [52]. Recently, evolutionary information extracted from frequency profiles have been combined with sequence-based kernels for protein remote homology detection. Profile-based protein representation has been used to extract evolutionary information into profiles [53]. Prioritizing and identifying reliable targets or pathways ahead for this traditional computational method may have a significant impact to the success of targets discovered.

Different data mining techniques ranging from text mining to emerging data mining approaches like chemo-genomic data mining and proteomic data mining are widely applied in the identification of potential drug targets [26]. Text/Literature mining is being broadly applied in identifications of diseases-associated entities and networks. Ozgur et al. [54] used new approach to predict gene-disease associations based on text mining and network analysis that has been tested on genes associated with prostate cancer. In this experiment, the gene called Epiregulin (EREG) is the only disease from the top 20 ranked genes where evidence has not been found about its association with prostate cancer but this is an important hypothesis which could be a good candidate for experimental study. The findings by Krauthammer *et al.* [55], is a good example for the application of text mining to identify disease- related networks. As the results were confirmed by experts in the field, this method performed well in predicting network nodes that match Alzheimer disease (AD) candidate genes. There are also mining tools which were developed for extracting interaction networks related to human diseases from literatures. For instance PolySearch [56] and GenCLip [57] could be mentioned as considerable efforts. There are some efforts in identifying drug targets of TB through data mining on datasets retrieved from literatures. Hologram QSAR (HQSAR) has been successfully applied to identify new anti-tubercular agents, targets and a minimum common bioactive substructure (MCBS) from the diverse chemical classes and within the particular chemical class that are known to be found in various drugs used for the current treatment of TB [58]. The dataset used for this analysis was extracted from literatures. Text mining has two stated limitations in spite of its promise in mining potentially useful knowledge from unstructured data [26]. One of these is term variation and term ambiguity of biomedical entities which may lead to erroneous relations between molecular biology and human diseases. The other is restricted access to the full text of papers. Since abstracts contain a high level summarization, mining from them could miss the detailed knowledge hidden in the main text body.

Microarray data mining has also been proven effective in identifying target genes associated with human diseases. Perry et al. [59] identified IGFBP3, which was selected and verified as a hypermethylation target of prostate cancer, through supervised classification technique. Ryu et al. [60] also strived to identify novel molecular signatures as therapeutic targets for aggressive melanoma with unsupervised clustering followed by supervised classification. Additionally, microarray data mining is promising in biomarker discovery. However, there are some challenges and limitations [26]. Results from microarray data mining has to be validated by follow up experiments because gene expression levels do not always correlate with protein levels. Microarray data from different labs are not always directly comparable. Moreover, different formats of data storage across databases have posed a great challenge.

The integrated data mining approach has been suggested as a solution to minimize the limitations of each data mining techniques [26]. For instance Li et al. [61] applied an integrated literature mining and gene expression analysis to model neuro-endocrine- immune (NEI) interactions. As it was indicated in the result, they identified numerous hub genes from the network that could be used as targets to inhibit tumor angiogenesis, such as tumor necrosis factor (TNF)-alpha, interleukin (IL)-1, -6 and vascular endothelial growth factor (VEGF).

Network-based approach is the other popular computational method which has been widely used in the identification of potential drug targets. Biological systems are a very complex assembly of a set of interacting molecules. Studying cellular level interactome network like protein-protein interactions has a great importance to understand the molecular underpinnings of life and for the like of assigning protein function which would be very useful for both basic research and drug development [46, 62, 63]. Network-based approach provides a powerful means to understand the complexity of biological systems and to reveal hidden relationships between drugs, genes, proteins and diseases. It has been successfully used for the identification of potential drug targets of complex diseases [64]. In recent years, various algorithms have been developed for these approaches. Some of the algorithms are based on local network properties and others focus on global network topology. Approaches that focus on global properties of the network consider global attributes like closeness/betweenness centrality measures and nodes with higher value of these attributes in the interaction network would be taken as initial drug target candidates.

There are few integrative computational methods which incorporates network-based approaches to identify therapeutic targets of TB. Raman et.al [65] successfully applied *in silico* target identification pipeline that includes protein-protein interactome network analysis, a flux balance analysis of the reactome, phenotype essentiality data, sequence analyses and a structural assessment. In this pipeline, a series of filters have been used where it has been started by identifying essential proteins through flux balance analysis and network analysis followed by comparative genomics with the host to prevent host toxicity. Then, non- similarity analysis with gut flora proteins and 'anti-targets' in the host, phrogentic profiling and gene expression analysis were carried out. Further, the resulted short listed proteins were filtered through mycobacterial persistence and drug resistance mechanisms. The resulted list of proteins were stated to be reliable drug targets since it has been shown through a thorough comparison with previously suggested targets from literatures. The method has been also suggested as gold standard approach for target identification and validation in TB. A new approach called crowed sourcing could be mentioned as another comprehensive method that has been used to identify novel drug target candidates of TB through an extensive re-annotation and analysis of system level protein-protein interaction network [66]. Mazandu and Nulder [67] also proposed potential targets of TB through integrated generation and analysis of large-scale data-driven functional network of the pathogen. They have generated functional interactome network by integrating functional genomic data. Then, drug targets were identified by analysing the topological properties of the resulted network.

There are considerable network based efforts that are particularly focused on trying to understand emergence of drug-resistance mechanisms from the wholistic approach [44, 68, 69]. In order to be successful in countering the problem of drug-resistance, it is important to understand about its emergence in the bacteria during drug treatment. The possible way to achieve this is through system level analysis of the proteins involved. In those methods, potential co-targets have been suggested to counter drug-resistance at the target identification phase. Co-target is a protein that needs to be simultaneously inhibited along with the intended primary target to prevent the emergence of resistance. Raman and Chandra [44] were successful in identifying the possible routes for the emergence of drug resistance in *Mycobacterium tuberculosis H37Rv* through *protein-protein* interactome analysis. The insights of those routes led them to identify proteins that are involved in the resistance- pathways. These proteins were proposed as co-targets to prevent the emergence of resistance. Chen et al. [69] also identified a list of co-targets of TB using random walk model on active drug-treated interactome network derived through heuristic search algorithm. The result showed that the active drug-treated networks are associated with the trigger of fatty acid metabolism, synthesis and nicotinamide adenine dinucleotide (NADH)-related processes. It has been stated that the result is in line with other experimental findings.

The common critics of network-based methods which are based on shortest paths is that they don't take all paths into account [46, 69]. The shortest path analysis probably provides a higher coverage than observed directly neighbours locally from the protein-protein interaction data but it only considers the shortest paths by ignoring the other paths that are longer and could be important for the communication in the interactome network.

Network flow approach has been successfully used to predict drug targets of prostate cancer from microarray data, disease genes and interactome networks [46]. Candidate proteins and disease genes were taken as an input and network flow approach were used to prioritize the candidate proteins based on their influence on disease genes. Through this procedure, the maximum flow and affected genes of 322 candidate proteins were identified. They carried out detail literature reviews for the top 22 candidate proteins. It has been stated that some of the proteins have been reported in the public literatures for validation. The strength of the method was that it doesn't suffer from the limitations of biasness of including only shortest paths. Based on their comparison with previous methods this approach has superior mean average precision and ranked the true drug targets higher.

Challenges and shortcomings of the current approaches show that there is a requirement for systematically integrated approaches to identify new therapeutic and preventive strategies. More specifically, it is crucial to focus the new strategies to drug target identification phase. It is a very important step because many of the failures in the expensive and time consuming drug discovery projects in the past are mainly related to ineffective target identification. Identification of reliable drug targets that have high potential of yielding clinical success within the efficacy–toxicity spectrum would be ideal solution to the challenges and threats caused by TB. The widespread of methods in target identification for TB has well documented drawbacks. Structure based approaches suffer from the limited availability of structures at least for the time being and the problem seems to continue until there will be significant progress in solving the structures of the majority of proteins or developing powerful computational techniques that can effectively predict structures. Network-based method has also limitations that range from poor quality of interactome network dataset to the method's biasness towards some nodes. Regardless of the stated limitations, it has been observed that network-based approaches have a potential to identify novel targets and reposition established targets successfully if they have been properly designed and executed.

In this study, protein-protein interactome network analysis has been used to identify potential drug targets of *Mycobacterium tuberculosis H37Rv.* The work flow of the analysis has been shown in Figure 1.3. The investigation can be broadly categorised into three main parts. The first one is to identify potential targets of the pathogen through network centrality measures, comparative genome analysis and maximum flow approaches. The network centrality measure was very helpful in identifying those proteins that found at the centre of interactome network. Targeting these proteins are believed to be effective in disrupting the network of the pathogen so that their inhibition could be successful in killing the bacteria or arresting its growth. The non-homologous analysis was carried out to filter out those proteins that are not homologous with the host for the purpose of avoiding host toxicity at the target identification phase of drug discovery process. Host toxicity is the main reason for the common preference of non homologous proteins as the primary choice for potential drug targets of various diseases. The other comparative genome analysis was to filter out those proteins that are essential for the survival and growth of the bacteria because the likelihood of a potential target to be successful increases if the protein is essential for the pathogen's survival and growth. A detail analysis was carried out on the list of potential drug targets of the pathogen to further prioritize them based on their influence on resistance genes through maximum flow approach. The belief in doing so is that the inhibition of a protein which has more influence on the resistance genes of the existing drugs is expected to disrupt the communication to these genes.

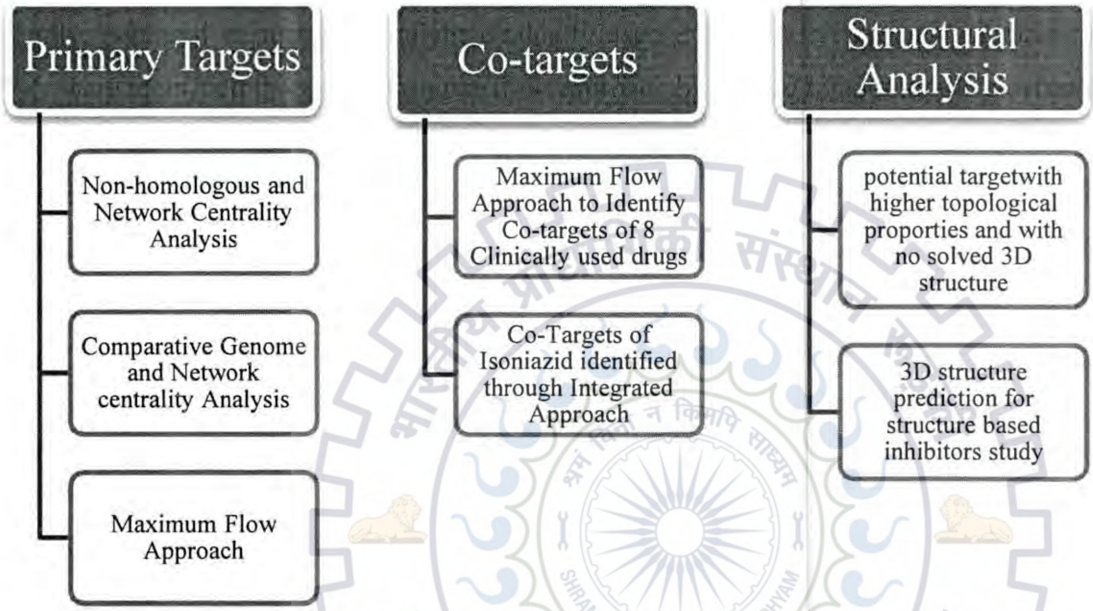

Primary Targets
Non-homologous and Network Centrality Analysis
Comparative Genome and Network centrality Analysis
Maximum Flow Approach
Co-targets
Maximum Flow Approach to Identify Co-targets of 8 Clinically used drugs
Co-Targets of Isoniazid identified through Integrated Approach
Structural Analysis
potential targetwith higher topological proporties and with no solved 3D structure
3D structure prediction for structure based inhibitors study

Figure 1.3 Work flow of the analysis

Secondly, potential co-targets for eight clinically used drugs of *Mycobacterium tuberculosis H37Rv* were identified through maximum flow approach by aiming to counter the emergence of drug-resistance. Proteins that are believed to be involved in the emergence of drug resistance were identified and prioritized based on their maximum flow value in the flow from drug targets of existing drugs to resistance genes by aiming the inhibition of a protein which has more influence on the resistance genes of the existing drugs is expected to disrupt the communication to these genes. Additionally, the systematic combination of comparative genome, network centrality measures and maximum flow approach has been used to identify and prioritise potential co-targets of Isoniazid.

Lastly, a potential target protein has been selected from the results of target and co-target identification analyses. The selected protein has higher network topological properties, non-homologous to human, essential for the survival and growth of the pathogen, with higher maximum flow value to resistance genes and without experimentally solved three-dimensional structure. *In silico* structural analysis has been carried out on the selected protein for structure-based inhibitor study.

1.3. Material and Methods

1.3.1 Biological Interactome Networks

Biological processes rely on the combined activities of bio-entities such as DNA, RNA, proteins, and metabolites [70]. The integration of activities among these entities can be described as an interaction network where the molecules are the vertices and their interaction as edges. The edges between two interacting bio-entities represent different types of relationships such as evolutionary relationship or the existence of a shared protein domain. It can also show that the two proteins belong to the same protein family or that two protein coding genes are co-expressed in an experiment [71]. Biological systems are complex and so are the corresponding networks by consisting several thousands of molecules and interactions. Edges could be either directed or undirected. In directed edges, there is a direction from the source node to target node. Directed edges can be useful for representing interactions like phosphorylation, where the source node represents the protein to phosphorylate, the edge is the phosphorylation process, and the target is the same protein in the phosphorylated state [72]. Commonly, undirected edges are used to represent interactions among proteins in protein-protein interaction networks. Generating and analyzing large scale interaction network will enhance our understanding of biological systems by providing numerous insights in system biology. The study of biological networks can follow descriptive approach like prioritizing nodes by their topological features to identify potential drug targets for development of pharmaceuticals or it could be predictive where it is usually applied to discover biologically significant facts [73]. The dataset for the construction of such networks is readily available in huge databases gathered from high throughput technologies ("omics"). Molecular level biological networks can be gene regulation networks, signal transduction networks, protein-protein interaction networks or metabolic networks [70]. The main focus here is on protein-protein interaction networks since the analyses in this study were based on protein-protein interaction network datasets.

Proteins are building blocks of living cells that control and mediate many of biological processes such as catalyzing, transporting and storing other molecules, controlling growth and development, providing mechanical strength, conferring immunity and transmitting signals [70]. In those biological processes, proteins don't operate alone rather they interact with other molecules such as low molecular weight compounds, lipids, nucleic acids, or other proteins. Protein-protein interactions are useful for the assembly of the cell's structural components like cytoskeleton. They are also highly important for various processes like transcription, splicing, and translation to cell cycle control, secretion, and the assembly of enzymatic complexes. Hence, protein-protein interaction networks are essential for many biological processes and compiling protein-protein interaction networks could be very helpful to enhance our understanding of the generic organization principles of functional cellular networks [74]. In order to have clear understanding about biological processes, we have to look at a protein from the context of its interacting partners.

One of the intended purposes of closely studying protein-protein interaction networks is to design potential therapeutic strategies for various human diseases since disease in many cases could be controlled by alterations of certain protein–protein interactions [70]. The completion and availability of complete genome sequences of several bacteria, virus and large eukaryotes have significant contribution in defining gene function at the morphological, biochemical, and physiological level. However, sequence information is not enough to have a clear understanding about the underlying principles of cellular systems mainly due to fact that many biological functions of plethora of predicted genes are experimentally uncharacterized. Protein-protein interaction networks fill this gap by providing system level insight about biological mechanisms and diseases processes which is much more than one at time studies of single components. Additionally, proteome interactome networks are used to annotate the function of unknown proteins through a principle called guilt by association [75]. The construction and analysis of genome level protein-protein interaction networks became a possibility due to the continuous and rapid advancements of highly parallelized and automated approaches.

There are three main methods for protein-protein interaction network identification; *in vitro*, *in vivo*, and *in silico* [76]. In *in vitro* methods, experiments are carried out in a control environment outside the organism. The techniques classified as *in vitro* include X-ray crystallography, NMR spectroscopy, tandem affinity purification, coimmunoprecipitation, protein arrays, protein fragment complementation and affinity chromatography, phage display. In *in vivo* methods, the protein-protein interactions are detected in the living organism. Yeast two-hybrid (Y2H, Y3H) and synthetic lethality are the two *in vivo* methods. *In silico* methods are computational techniques where the interaction network is identified through computer simulations. *In silco* techniques include sequence-based approaches, gene fusion, structure-based approaches, *in silico* two-hybrid, chromosome proximity, mirror tree, phylogenetic tree, and gene expression-based approaches.

Currently, computational methods are capable of generating protein-protein interaction networks from the vast amount of diverse biological data such as genome sequences and protein structures [72]. Various microarray experiments and the resulted gene expression data stored in public biomedicine databases further allow us to infer interactions. However, careful consideration and caution has to be taken in using this abundantly available interaction data because information could be misleading due to the presence of false positives and false negatives. Usually, biological interaction networks derived from a single omic approach only provide crude gene or protein function [75]. Thus, it has been recommended to use integrated interaction datasets obtained through various computational and experimental methods. The integration definitely improves functional annotations and can be used to formulate biological hypotheses.

1.3.2 Statistical Properties of Interaction Network

As it has been stated, intricate connectivity of cellular systems can be represented by complex networks. In these networks, vertices represent interacting entities and edges represent various forms of interactions between them. For instance, in transcriptional network, genes or proteins are represented by vertices and regulatory interactions are represented by edges. In some cases, complex networks are used to represent more abstract processes like a vertex may represents different configurational states of a protein and edge represents transitions between them. These complex biological networks can be systematically characterized by their network statistical properties such as degree, degree distribution, characteristic path length and clustering coefficient. They are used to describe the underlying design principles, unknown organizing principles and the functional organization of cellular systems [70]. Insights about evolution of interacting molecules, the influence of organization on their function and dynamic responses could also be obtained from network topological properties [77].

Let V be a set of nodes representing biological molecules like protein or gene and E be a set of edges representing interactions between them. Then, the biological network is formally represented by the network G = (V, E). The size is denoted as n=|V| and m=|E|. In undirected network, edges don't have directions whereas in directed graphs, edges contain directional information. The network could also be either weighted or non-weighted graph. If it is a weighted network there would be a scalar value associated with each edge that is used to quantify interaction strength, cost or flow.

Degree k_i is one of the most basic properties of a vertex n_i in an interaction network, which is defined as the number of edges incident to the vertex. If a molecule interacts with various molecules it would have a high node degree. For instance, ATP usually interacts with many proteins so it has high node degree [72]. For directed network, degree is distinguished as *input degree* which is represented as k_i^{in} and *output degree* represented as k_i^{out}.

Degree distribution p (k) of a network is important characteristics of network topology which is defined as the measure of the proportion of nodes in the network having degree *k* [70]. The degree distribution of numerous empirical networks were observed to follow power law stated as $p(k) \sim k^{-\gamma}$, where γ is a representation of degree exponent which is commonly in the range of $2 < \gamma < 3$ [70, 77]. If the given network exhibits the stated power law distribution then the network is said to be scale-free network [72]. A scale-free network consists of a high diversity of node degrees where there exist large numbers of nodes with small degree and very few numbers of nodes with high degree. The few nodes that have high degree are called hubs. Many biological networks exhibit the property of scale-free [72]. If a given network is scale-free, it is resistance to random attacks where random removal of nodes is likely to take out small degree nodes [78]. This means the removal of randomly selected nodes will not have significant impact on the connectivity of the network since the few highly connected nodes preserve the connectivity. However, the network could be disconnected by targeted attacks on the hub nodes. In a given network with n vertices, two nodes n_i and n_j are said to be connected if there exist a path between them. A path is an alternative sequence of distinct edges that needs to be traversed from one vertex to another and it could be a representation of a sequence of interactions like a molecular binding to a receptor followed by transcription factor binding to DNA. A path may not be unique since there could be more than one path even with the same path length between any two vertices. The distance d_{ij} between two vertices is the length of the shortest path between them. If the network is a directed or disconnected graph there may not be a path between any two arbitrarily taken vertices then the distance between these vertices is set to infinite. In the case of weighted graph, the distance or path is not straight forward since it has to accommodate scalar information called weight. For instance, the shortest distance may not necessarily be the cheapest.

The average shortest path which is also alternatively called characteristic path length of a given network is computed by averaging all of the shortest paths between any two pairs of nodes. It is an important property of a network which is used to describe how quickly information is navigated throughout the network [72]. Unsurprisingly, most networks have very small characteristic path length that has been well described by the theory of six degree of separation. Let us have a network consisting of people in the planet as vertices and their interactions as edges. Then, there are very few numbers of intermediate friends (six) who separates any two persons in the network and this is called six degree of separation. Networks that exhibit such kind of property are called small world networks. More specifically, a given network can be called small world if the growth of its characteristic path length is approximately equal to the growth of logarithm of the network size. Many biological networks such as metabolic and protein-protein interaction networks are small world networks [75].

Another important property of the network which shows local cohesiveness is the clustering coefficient C [70]. It is a measure of the probability that two nodes with a common neighbour to be connected. It is an indicator of the internal structure of the network. In undirected network, for a given node n_i with k_i neighbours, there exist $E_{\max} = \frac{k_i(k_i-1)}{2}$ possible edges between the neighbors. Clustering coefficient c_i of vertex n_i is then given as the ratio of the actual number of edges E_i between the neighbors to the maximal number $E_{\max}$;

$$c_i = \frac{2E}{k_i(k_i-1)} \tag{1}$$

The global or mean clustering coefficient C_i of the network is the average cluster coefficient of all vertices.

1.3.3 Network Centrality Measures

Network centrality measures are used to numerically characterize the importance of nodes in a specified network. In a protein-protein interaction network, network centrality measures can be used to indicate the importance of proteins in the system, and their contribution to the functioning of the system. Thus, they can be helpful in assessing the topological significance of these proteins within the network and quantifying the structural properties of the produced functional network. There are many network centrality measures but our discussion is only focused on four of the more widely used centrality measures which are degree, closeness, betweenness and eigenvector.

Let G = (V, E) be an undirected graph and A is $n \times n$ adjacency matrix. The degree centrality is defined as:

$$ki = \sum_{j=1}^{n} A_{ij} \tag{2}$$

$$\text{where } A_{ij} = \begin{cases} 1 \text{ if node } i \text{ is connecetd to node } j \\ 0 \qquad\qquad\qquad\qquad \text{otherewise} \end{cases}$$

The degree centrality measure ranks an individual node in the network based on its connectivity. In a biological network, degree is an indicator of influence of a molecule on the biological processes. For instance a protein with higher degree in protein-protein interactome network tends to contribute to several processes, and potentially be a key protein in the functioning of the system.

The closeness centrality of a node measures how close a node is to other nodes in the network. The smaller the total distance of a node to other nodes, the higher its closeness is. We calculate closeness centrality measure for a node by inverting the sum of the distances from it to other nodes in the network.

The closeness centrality is defined as:

$$C_{clo}(s) = \frac{1}{\sum \mathrm{dist(s,t)}} \qquad (3)$$

The closeness measure is high for a node that is central since it has a shorter distance on average to other nodes.

The betweenness centrality of a node *v* in a network is a metric that expresses the influence of *v* relative to other nodes within the network. It is based on the proportion of shortest paths between other nodes passing through the node in consideration. For instance, if the node is a protein it shows its importance for the transmission of information between other proteins in the network. This metric provides an indication of the number of pair-wise proteins connected indirectly by the protein target through their direct functional connections.

The betweenness, *B* (*v*), of a node *v* is given by:

$$B(v) = \sum_{s \neq v \neq t} \frac{\sigma st\ (v)}{\sigma st} \qquad (4)$$

where *σst*(*v*) is the number of shortest paths from protein *s* to protein *t* passing through *v* and *σst* the number of shortest paths from *s* to *t* in the functional network.

In a protein-protein interaction network, proteins with higher rank of betweenness are expected to ensure the connectivity between proteins in the functional network and are able to bridge or disconnect connected components. Such proteins are hubs, referring to proteins that are highly connected and serve to hold together a large number of proteins with low degree, thus integrating all proteins in a given connected component into a unified complex system. These proteins are of utmost importance for the integrity and the robustness of the system and are responsible for the small world property since connections between proteins are relatively short via these hubs.

While we are computing degree centrality all connections between nodes are equally important but in reality some of the relationships (connections) between neighbours are clearly more valuable than others. This notion is defined as 'prestige' in social networks. Intuitively, the prestige of a person does not only depend on the number of acquaintances he has, but also how prestigious his acquaintances are. A node in a network is more central if it is connected to many central nodes. The centrality x_i of node i is proportional to the sum of the centralities of its neighbours [79]:

$$\mathrm{x_i} = \lambda^{-1} \sum_{j=1}^{n} Aijxj \qquad (5)$$

Let's represent the centralities of the nodes as a vector x=(x1,x2,...,xn) and rewrite equation (5) in matrix form.

$$\lambda \mathrm{x} = \mathrm{Ax} \qquad (6)$$

Here, x is an eigenvector of the adjacency matrix A with eigenvalue λ. By Perron–Frobenius theorem, there is only one eigenvector x with all centrality values non-negative and this is the unique eigenvector that corresponds to the largest eigenvalue λ [79]. Eigenvector centrality assigns each node a centrality that not only depends on the quantity of its connections, but also on their qualities.

In a protein-protein interaction network, eigenvector centrality metric assigns a relative weight to all proteins in the network based on the fact that functional connections are not equally important and functional connections to more influential proteins will impact more on the contribution of the protein than functional connections to less influential proteins. Thus proteins with a high number of functional interactions are important, but a protein with a small number of high-quality functional connections may contribute more to the survival of the organism than one with a large number of low-quality functional connections.

1.3.4 Network Flow Approaches

Network flow approach is very widely accepted and well established stream that has a long tradition of being applicable in many fields including applied mathematics, computer science, engineering, management, and operations research [80]. Recently, it has been effectively used in the prediction of potential drug targets of prostate cancer [46]. The problem of network flow can be broadly categorised into three. These are minimum cost flow, shortest path and maximum flow problems. The general definitions and applications of all the three network flow problems are introduce but a more emphasis was given to a maximum flow problem since it has been used to identify potential drug targets and co-targets for *Mycobacterium tuberculosis H37Rv* in this study.

Minimum Cost Flow Problem: It is one of the fundamental network flow problems where the objective is to identify the possible minimum cost paths for the flow from one or more nodes in the network to one or more other nodes [80]. Some of the popular applications of minimum cost flow problems are like the flow of raw materials and intermediate goods in a production line, automobiles routing through an urban street network and the telephone system's routing of calls.

Let G = (V, E) be a directed graph where V denotes a set of *n* nodes and E denotes a set of *m* directed edges. Each edge $(i,j) \in E$ has an associated cost c_{ij} that represents the cost per unit flow in the edge. Additionally, each edge $(i,j) \in E$ has a capacity U_{ij} which represents the maximum amount that can possibly flow on the edge. The supply/demand in each node i is represented by *b(i)*. The node i is said to be a supply node if $b(i) > 0$ where as if $b(i) < 0$ it is a demand node with a demand of $-\,b(i)$. But node *i* is a *transshipment node* if $b(i) = 0$.The flow on the edge $(i,j) \in E$ is denoted by x_{ij}.

Then, the minimum cost problem is formulated as:

$$\text{Minimize} \quad \sum_{(i,j)\in E} c_{ij}x_{ij} \tag{7}$$

$$\text{Subject to} \quad x_{ij} \le U_{ij} \tag{9}$$

$$\sum_{\{j:(i,j)\in E\}} x_{ij} - \sum_{\{i:(j,i)\}} x_{ji} = b(i) \ \text{ for all } i \in V \tag{8}$$

$$\text{where } \sum_{i}^{n} b(i) = 0$$

Alternatively, minimum cost problem can be represented in the following matrix form:

$$\text{Minimize} \quad cx \tag{10}$$

$$\text{Subjected to} \quad Nx = b \tag{11}$$

where *N* is $n \times m$ matrix

The flow x_{ij} has to always satisfy the capacity constraint (8) and any node has to satisfy mass balance constraint (9) which states that the difference between outflow and inflow must be equal to the supply/demand of the node under consideration. The outflow is commonly greater than inflow in supply nodes and outflow is less than inflow in demand nodes. But if the node is *transshipment node* the outflow is equal to the inflow.

Shortest path problem: of all the three network flow problems, shortest path problem is probably the simplest where the objective is to find the shortest path from a distinguished *source node s* to another specified *sink node t*. It has been widely applied in many problem domains like cash flow management, project scheduling, equipment replacement, message routing and traffic flow [80]. In some cases, the shortest path problem is intermingled with minimum cost problem. In the minimum cost flow problem, let *b(s)* = 1, *b (t)* = - 1, and b (i) = 0 for all other nodes, then the solution to the problem will send 1 unit of flow from the source node *s* to sink node *t* through the shortest path. The shortest path problem is also used to model situations in sending flow from a single source node to a single sink node.

Maximum flow problem: the maximum flow problem is a little bit complementary to shortest path problem where in the later case; the solution of a problem is modeled in terms of cost in contrast to maximum flow problem that models situations in terms of flow. The flows are restricted by flow bounds [80]. The objective in maximum flow problem is to find the possible maximum flow from a distinguished source node *s* to another distinguished sink node *t*. It is widely applied in identifying maximum flows of petroleum products in a pipeline network, messages in a telecommunication network, cars in a road network and electricity in an electrical network. The maximum flow problem can be formulated into minimum cost problem. Let $b(i) = 0$ for all $i \in N, c_{ij} = 0$ for all $(i,j) \in E$ and assume (t,s) be an additional edge with cost $c_{ts} = -1$ and capacity $U_{ts} = \infty$. Then, the objective of minimum cost flow is to maximize the flow on edge (t,s). The flow on edge (t,s) is only possible from node *s* to node *t* via intermediate edges of E. Thus, the formulated minimum cost problem will maximize the maximum flow from source node *s* to sink node *t*. Maximum flow problem has been discussed more in the following section.

1.3.5 Maximum Flow Approach

As it has been stated, maximum flow problem is a classical combinatorial problem that has a wide range of applications in a number of areas [80, 81]. The purpose in maximum flow approach is loud and clear which is to send the maximum possible flow between two special nodes called source node s and sink node t in a capacitated directed network without violating the capacity constraint.

In a capacitated directed network G = (V, E), let us assume that u_{ij} is a nonnegative capacity associated with each edge $(i,j) \in E$. Suppose source node s and sink node t are two distinguished special nodes. The aim is to find the maximum flow from s to t by satisfying both capacity and flow conservation constraints. Then the problem of maximum flow could be defined as follows:

$$\text{Maximize } v \qquad (12)$$

$$\text{Subjected to } \sum_{\{j:(i,j)\in E\}} x_{ij} - \sum_{\{j:(j,i)\in E\}} x_{ji} = \begin{cases} v & for\ i = s \\ 0\ for\ all\ i \in N - \{s,t\} & \\ -v & for\ i = t \end{cases} \qquad (13)$$

$$\text{where } x_{ij} \leq u_{ij} \text{ for each } (i,j) \in E \qquad (14)$$

Residual network is a network derived from the original network of interest and very important in the implementation of maximum flow problems using various algorithms. The residual capacity r_{ij} of an edge $(i,j) \in E$ in residual graph is the maximum additional flow that can be increased on the edge (i,j) without violating the capacity constraint. It is calculated as $r_{ij} = u_{ij} - x_{ij} + x_{ji}$. Therefore residual network is constructed from residual edges with positive capacity.

There are various algorithms for the implementation of classical maximum flow problem such as Dantzig's network simplex method [82, 83], Ford and Fulkerson's augmenting path method [84], Dinitz's blocking flow method [85] and the famous push-relabel method of Goldberg and Tarjan [86,87]. Generally, the implementation algorithms can be categorized into two types [80]:

1. ***Augmenting path algorithms:*** are algorithms where the flow is incrementally augmented from the source node s to sink node t through different paths without violating flow conservation constraint at every node of the network other than the distinguished source and sink nodes.
2. ***Pre-flow push algorithms:*** works by incrementally pushing excess flow value from active nodes. The flow could be either forward from the active node v towards the sink node t or backward towards the source node s.

Generic augmenting path algorithm is the most simple and intuitive algorithm for solving many maximum flow problems [80]. Augmenting path is defined as a directed path from the source node *s* to sink node *t*. The residual capacity of an augmenting path is the minimum residual capacity of any edge in the path. Residual capacity of an augmenting path is always positive and generic augmenting path algorithm works by the principle of iteratively sending additional flow from source node *s* to sink node *t* through the augmenting paths until the network contains no such paths. The augmenting path algorithm is described as follow [80]:

Augmenting path algorithm:

```
begin
    x: = 0;
    while G(x) contains a directed path from node s to node t do
    begin
        identify an augmenting path P from source node s to sink node t;   $\delta := min\{r_{ij}: (i,j) \in P\}$;
        augment $\delta$ units of flow along P and update G(x);
    end;
end;
```

Maximum flow algorithms which are based on generic augmenting path have two worth mentioning advantages; their power to solve any maximum flow problem with integral capacity data and their ability in providing us a constructive tool for establishing the fundamental max-flow min-cut theorem [80]. Because of these, they are helpful in the development of many combinatorial applications for network flow theory. However, they have two major computational limitations. The first one is their worst case complexity *($O(nmU)$)* that undermines their application for maximum flow problems with large capacity. The other is that the algorithms probably converge to a non optimal solution in the case of irrational capacity data.

Pre flow-push algorithms are based on the manipulation of pre-flow *f* that satisfies capacity constraints. Besides the pre-flow *f*, these algorithms maintain distance level *d*. In recent times, they have been emerged as better and powerful methods in terms of both theory and practice for solving maximum flow problems. The algorithm works by iteratively pushing flows on individual edges in contrast to augmenting path which uses augmenting paths. Because of that pre-flow push algorithm doesn't necessarily satisfy mass balance constraint [80, 81].

Pre-flow *f* of each edge $(i,j) \in E$ has to be maintained at the intermediate stages of the execution of the algorithm. The excess *e* of a node $i \in V$ is formulated as;

$$e(i) = \sum_{\{j:(j,i)\in E\}} x_{ji} - \sum_{\{j:(i,j)\in E\}} x_{ij} \begin{cases} \geq 0 \; for\; all\; i \in V - \{s\} \\ \leq 0 \qquad\quad for\; i = s \end{cases} \tag{15}$$

A node v is said to be active if its excess $e(v) > 0\; for\; all\; v - \{s,t\}$ and distance $d(v) < n$.

In pre-flow push algorithms, an active node is being selected and then, its excess would be pushed towards its neighbors. The excess is pushed first to the neighbor that is closest to the sink node t since the ultimate objective is maximizing the flow towards the sink node. In other words, the edge which carries excess from active nodes has to be an admissible edge. An edge (i,j) is said to be admissible if $d(i) = d(j) + 1$. If there is an active node of interest with no admissible edges, its distance label will be increased to make it admissible. This is called distance relabeling and it can be implemented through either global relabeling or gap relabeling heuristics. The algorithm terminates when there are no more active nodes in the network. The subroutines of pre-flow push algorithm are shown as follows but there exist various implementations [80];

```
procedure preprocess;
begin
        $x = 0$;
        compute the exact distance labels $d(i)$;
        $x_{sj} = u_{sj}$ for all edge $(s,j) \in E$;
        $d(s) = n$;
end;

procedure push-relabel(i);
begin
        if the network contains an admissible arc (i, j) then
                push $\delta = min\{e(i), r_{ij}\}$ units of flow from node i to node j
        else replace d(i) by min $\{d(j) + 1: (i,j) \in E \; and \; r_{ij} > 0\}$;
end;
```

1.3.6 Datasets and Tools

In this study, various software packages, program tools, materials and datasets have been used as source of data, to analyse it and for visualisation of results. Lists of these resources are provided as follows;

1. **Software Packages and Program Tools**
   - Basic Local Alignment Search Tool (BLASTp) [88]
   - Cytoscape 3.0.2 [89]

- CytoNCA [90]
- C++
- Modeller 9.11[91]
- RCSB Protein Data Bank (PDB) [92]
- Pymol [93]
- WinCoot [94]

2. ***Mycobacterium tuberculosis H37Rv* Datasets**
   - Complete genome from Tuberculosis database [28,95]
   - Protein-Protein Interactions from STRING database [96]
   - Non-human homologous proteins from Drug Target Protein Database [97]
   - Essential genes from Database of Essential Genes (DEG) [98-100]
   - List of curated drug resistance genes from literatures [44,68]
   - List of validated drug targets from literatures [68]
   - List of proteins that interact with the host from literatures [101,102]
   - Drug targets from Drug Target Protein Database [97]
   - UniProt target list from UniProt [103]
   - TDR validated targets [104]

1.4. Summary of the Thesis

In this work, systematically integrated protein-protein interaction network analyses have been carried out to identify potential drug targets and co-targets of Mycobacterium tuberculosis H37Rv. These have been done by aiming to counter the problems of TB at the first phase of drug discovery process. *In silco* molecular modelling and structure analysis has been carried out for protein translocase subunit SecY (Rv0732). Further, future perspectives have also been incorporated to give insight about future investigations regarding to drug target discovery of the pathogen. With this perspective, chapters of the thesis have been summarized as follows:

Chapter 1: A brief introduction to drug discovery has been given which includes historical perspective, paradigms in the modern drug discovery process, roles of computational techniques in the drug discovery pipelines and existing challenges. A more specific discussion about target identification and its impact on the failures of previous drug discovery has been incorporated. It also includes related basic concepts. A thorough review of related literatures has been carried out on drug target identification using computational techniques with a more focus on TB. The research problem has been formulated based on the observed gaps. Subsequently, detailed discussions have been carried out about the materials, methods and tools which are used in these analyses.

Chapter 2: Most central and non-homologous proteins have been identified and proposed as a list of potential primary drug targets of *Mycobacterium tuberculosis H37Rv*. The analysis has been started through the construction of pathogen's protein-protein interaction network using a dataset retrieved from STRING database. Then, the four centrality measures such as degree, closeness, betweenness and eigenvector have been used to identify the most central proteins from the constructed protein-protein interaction network by hypothesizing that these proteins would be important to alter the function of the network. Proteins that are non-homologous with the host are always the primary preferences as drug targets especially for host–parasite diseases like TB. Thus, the list of proteins resulted from network centrality analysis have been filtered out using a dataset obtained from Drug Target Protein Database to identify those proteins which are non-homologous to human. For the purpose of validation, the final list of proteins has been compared with the previously reported targets. Furthermore, an assessment about the structural coverage of the proposed targets has been carried out.

Chapter 3: A systematically integrated comparative genome and network centrality analysis has been used to identify more reliable drug targets of *Mycobacterium tuberculosis H37Rv*. The identified targets are believed to be more reliable because of the use of integrated approach. Initially, the complete genome sequence dataset of protein coding genes of the pathogen has been retrieved from Tuberculosis Database. Then, comparative genome analysis has been carried out on the retrieved genes against Database of Essential Genes to identify genes that are essential for the survival and growth of the pathogen. The resulted essential genes have been filtered out by blasting against non-redundant database with an e-value threshold cut off 0.005 and restricting the organism to H. sapiens to identify genes that are non homologous to human. The protein-protein interaction network of the pathogen has been constructed using an interaction dataset from STRING and the filtered list of essential and non human homologous proteins have been prioritized using network centrality measures. Those proteins that found at the center of gravity of the interactome network have been proposed as final prioritized list of potential drug targets of the pathogen. Further assessments including, structural coverage, comparison with known and proposed drug targets, comparison with list of proteins that interact with the host have been carried out.

Chapter 4: The emergence of drug-resistance varieties in TB is the main challenge where the current treatment strategies are less effective, more expensive, take longer duration and with numerous side effects. This indicates the requirements of new therapeutic and preventive strategies to counter the problem. Potential drug targets of *Mycobacterium tuberculosis H37Rv,* obtained from previous analysis, have been further prioritized based on their influence to resistance genes by aiming the inhibition of a protein which has more influence to resistance genes of the existing drugs will disrupt the communication to these genes. For this analysis, a weighted protein-protein interaction network of the pathogen has been constructed using a dataset retrieved from STRING. The potential drug targets from the previous analysis and resistance genes from literatures have been taken as inputs. Then, the potential drug targets have been prioritized based on their maximum flow value to resistance genes. The method can be used as additional druggablity assessment criteria for drug resistance diseases like TB.

Chapter 5: The main strategy to counter the problem of resistance has to be started by enhancing our limited system level knowledge about the possible routes and causes of resistance. With this perspective, this analysis has been carried out to identify potential co-targets for eight clinically used drugs of TB through maximum flow approach to prevent the emergence of drug resistance. The analysis has been started by constructing drug-specific protein-protein interaction networks of the pathogen. The validated drug targets of drugs used in the current regime of TB treatment and resistance genes have been taken as inputs. Then, the maximum flow values of proteins in the flow from validated drug targets to resistance genes have been identified. Proteins have been prioritized based on their maximum flow value. Subsequently, filters such as non-homologous assessment to avoid host toxicity, identification of proteins that interact with the host and essentiality analysis have been carried out. The final refined lists of proteins have been reported as potential co-targets for each drugs of the pathogen. It is believed that these proteins have strong involvement in the emergence of drug resistance by being used as mediators of information from drug targets to resistance machineries. Thus, targeting them with systematic combination of existing drugs is believed to be effective to prevent the emergence of drug resistance.

Chapter 6: The availability of structural information of a specified drug target is one of the druggablity criteria. However, many proteins do not have experimentally solved structure in spite of the efforts of structural genomics projects. Homology (comparative) modeling techniques have been widely used to minimize this problem. Thus, structural analysis on a selected potential drug target of *Mycobacterium tuberculosis H37Rv* has been carried out in this analysis. From the previous analyses, protein translocase subunit SecY (Rv0732) has been selected since it is highly ranked potential drug target without solved three-dimensional structure. An *in silco* structural analysis has been carried out to get descriptive three-dimensional structure. The active site has been identified for protein-ligand or protein-inhibitor binding. The identification and characterization of the binding site is a vital step in the process of structure-based drug design.

Chapter 7: In this chapter, the main conclusions are drawn from the computational analyses to identify and prioritize potential primary drug targets and co-targets of *Mycobacterium tuberculosis H37Rv.* The future perspectives of drug target discovery in TB have been indicated.

CHAPTER TWO

Potential Non-Homologous Protein Targets of *Mycobacterium Tuberculosis H37Rv* Identified from Protein-Protein Interaction Network

2.1. Introduction

According to WHO reports and estimates, there is a slow decline of deaths caused by TB [29]. However, its universal burden is still significantly high. Especially if we consider the fact that most of TB deaths are preventable, the global mortality rate from this disease is unacceptable. One of the challenges is the existence of enhanced susceptibility to TB in HIV-infected people. The other is emergence and spread of various forms of drug-resistance TB not only in developing countries but also in industrialized nations [105, 106]. The TB reports show that there are few promising candidate drugs of TB at the late stage of development phase such as diarylquinoline TMC207, nitroimidazopyran PA-824, nitroimidazo-oxazole Delamanid (OPC-67683), oxazolidinone PNU-100480, ethylene diamine SQ-109, and pyrrole derivative LL3858 [107]. These candidate drugs are under phase 1 to phase 3 of clinical trials. Nevertheless, the TB drug development specifically for those truly active against dormant and persistent types of *tubercle bacilli* is unsuccessful and slow. Thus, there is an urgent need of novel anti-tuberculosis drug targets from which new drugs can be developed that can act on establishment of mycobacterial dormancy in the host's macrophages.

It is possible to identify potential therapeutic target of TB through a range of computational approaches using the readily available biological data and information provided that there would be a follow up experimental validation. There are considerable efforts that could be mentioned with this regard where computational techniques have been applied to identify and validate new drug targets. One of the common approaches is the use of structural information to predict if a protein can be a drug target [46]. As it has been stated previously, the limited availability of protein three-dimensional structures has been the main challenge for this method. There are several attempts to predict drug-target associations by combining drug-drug and gene-gene similarity measurement where the chemical structure of the drugs and sequence information of the target proteins were used to be the features to learn the classifier based on the target proteins as gold-standard positive dataset [47]. Once again the challenge here is that the two-dimensional structure may not necessarily be translated into the corresponding three-dimensional structure.
Computational methods have been used to identify potential protein targets of *mycobacterium tuberculosis* through the analysis of protein-protein interaction network. Studying genomic scale

protein-protein interaction networks have indispensable significance to understand the networks in living cells and for the like of assigning protein function which would be very useful for both basic research and drug development [62, 63]. Proteins normally have to interact with each other to do their normal job. This means most of their functions that are important for the proper functioning of life are associated with protein-protein interactions. The functions that are affected by protein-protein interactions include some proteins that are used as inhibitors of enzymes, proteins that are directed to the correct compartments of cells by binding to other proteins, protein messengers that bind to protein receptors on the outer surface of cell membranes to send signals between cells, interactions between different protein subunits which are the basis of allosteric changes in oligomers and very large-scale movements in organisms like muscle contraction that are triggered by protein–protein interactions. All cellular processes are dependent on precisely orchestrated interactions between proteins. Protein-protein interaction network has been used as an integrated part in identification of non-homologous proteins (enzymes) as potential drug target for *Mycobacterium tuberculosis* [108]. In the study, the protein-protein interaction network analysis was carried out to identify the most potential metabolic functional associations among all identified choke point proteins and to find out the functional association of high interacting metabolic proteins to pathogenesis causing proteins.

The analysis of the protein-protein interaction network and identification of most central proteins that are engaged in the process is believed to have a great importance because proteins interact with each other to accomplish most of the process in living cell. The interaction networks provide a powerful means to understand the complexity of biological systems and to reveal hidden relationships. In this analysis, the protein-protein interaction network of *Mycobacterium tuberculosis* have been constructed and analysed to identify most central non-human homologous proteins as potential drug targets. The four network centrality measures; degree, closeness, betweenness and eigenvector have been used for the identification of the potential drug targets.

2.2. Materials and Method

The protein-protein interaction network of *Mycobacterium tuberculosis H37Rv* was constructed by using a dataset from STRING database. STRING is a database and web resource dedicated to protein–protein interactions, including both physical and functional interactions [96]. It weights and integrates information from numerous sources, including experimental repositories, computational prediction methods and public text collections. Hence, it acts as a meta-database that maps all interaction evidence onto a common set of genomes and proteins. Due to the low quality of available datasets for *Mycobacterium tuberculosis H37Rv*, the interactome network may contain false positives as well as false negatives. To reduce its impact only 'high-confidence' and 'medium-confidence' data are being used. The statistical properties of the generated proteome network were characterized by different measures such as degree distribution, characteristic path length and clustering coefficient to understand the general functional organization of interacting proteins. Network visualization and analysis was carried out with Cytoscape 3.0.2 [89]. It is an open source software platform designed for visualizing complex networks and integrating these with any type of attribute data.

Four network centrality measures namely degree, closeness, betweenness and eigenvector have been used to rank proteins in the proteome interaction network. The network centrality measure of a protein is a way of numerically characterizing the protein's importance and its contribution to the functioning of the system in the protein-protein interaction network. This means network centrality measures are helpful for assessing the topological significance of proteins within the network by quantifying the structural properties of the functional network. The four centrality measure values of each protein in the constructed protein-protein interaction network of *Mycobacterium tuberculosis* were computed using CytoNCA. It is a plug-in of Cytoscape [90].

Further, the most central proteins which were identified and prioritized by the stated centrality measures were refined to identify non-human homologous proteins since in host–parasite diseases like tuberculosis, non-homologous proteins (enzymes) as drug target are of first preference. For this purpose, a list of non-human homologous proteins obtained from Drug Target Protein Database has been used [97].

2.3. Results and Discussion

2.3.1 Interactome network

A proteome-scale interaction network of proteins in *Mycobacterium tuberculosis* was derived from the STRING database [96]. In STRING, a confidence score is assigned to each identified protein-protein association, derived by benchmarking the performance of the predictions against a common reference set of trusted, true associations. A higher score is assigned when an association is supported by several types of evidence, thus expressing increased confidence. For each evidence source, functional interaction scores are categorized into three different confidence levels namely; low, medium, and high confidence. Note that for a given data source, all interactions whose scores are strictly less than 0.4 are considered as low confidence, scores ranging from 0.4 to 0.7 are classified as medium confidence, and scores greater than 0.7 yield high confidence [109]. A recent comprehensive study was carried out on the evaluation of the two main protein-protein interaction datasets of *Mycobacterium tuberculosis H37Rv*; B2H and STRING and it indicated that these datasets are of low quality which seems to contain a significant amount of false positives as well as false negatives [110]. They also stated that a subset comprising protein-protein interactions with higher scores in STRING dataset is more reliable. To minimise this problem, all interactions tagged as 'low-confidence' in the STRING database have been eliminated from this study.

Table 2.1 Network statistics

Parameter	**Value**
Number of nodes (n)	3956
Connected components	8
Network diameter	10
Average number of neighbours	32.556
Network density	0.008
Network heterogeneity	0.942
Shortest paths	15519694(99%)
Characteristic path length	3.096
Clustering coefficient	0.294

The resulted *Mycobacterium tuberculosis* protein-protein interaction network consists of 3956 distinct proteins as nodes and 128854 edges of interactions among these proteins. Statistical properties of the generated network have been shown in Table 2.1 to describe its essential properties. The characteristic path length of the network, which is the average distance between all pairs of nodes, is smaller than $log(n)$. This implies that the *Mycobacterium tuberculosis H37Rv* proteome interaction network has "small world property" [72]. This property provides an idea about the network's navigability by indicating how fast information can be communicated in the system irrespective of the number of nodes. Thus, from this small world property of the network, we can infer that there are efficient communications of information where the transmission of biological information from a given protein to others can be achieved through only a few steps. This means one protein can have an influence on another with only a small number of intermediate reactions. With this property the organism could obtain an evolutionary advantage of being able to respond quickly to perturbations in the environment and to exhibit a qualitative change of behaviour in response to these perturbations [67].The shortest path length distribution between pair-wise protein interactions has been shown in Figure 2.1.

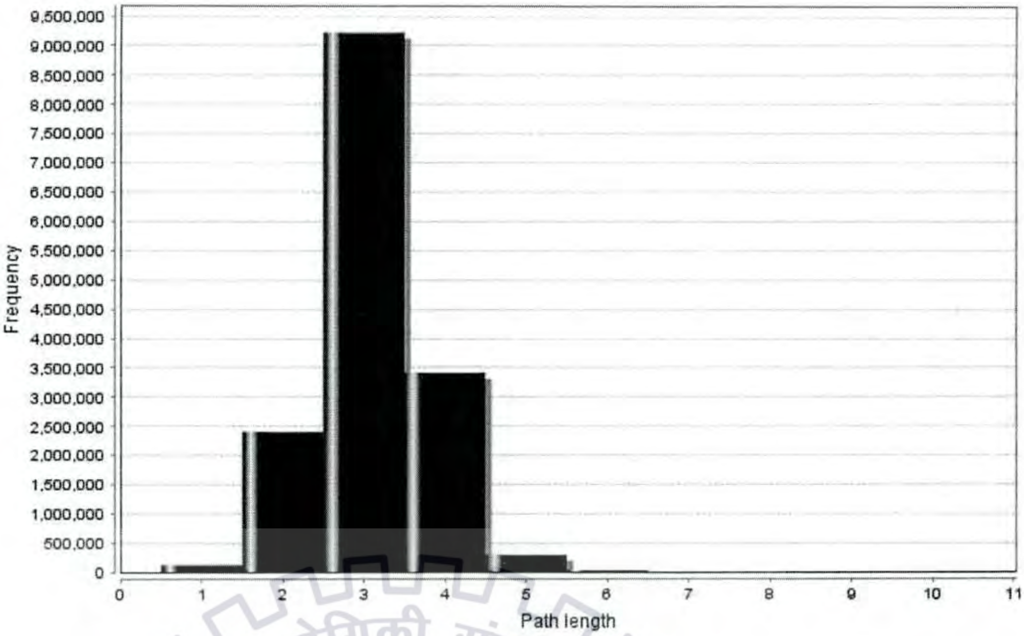


Figure 2.1 Shortest path length distributions

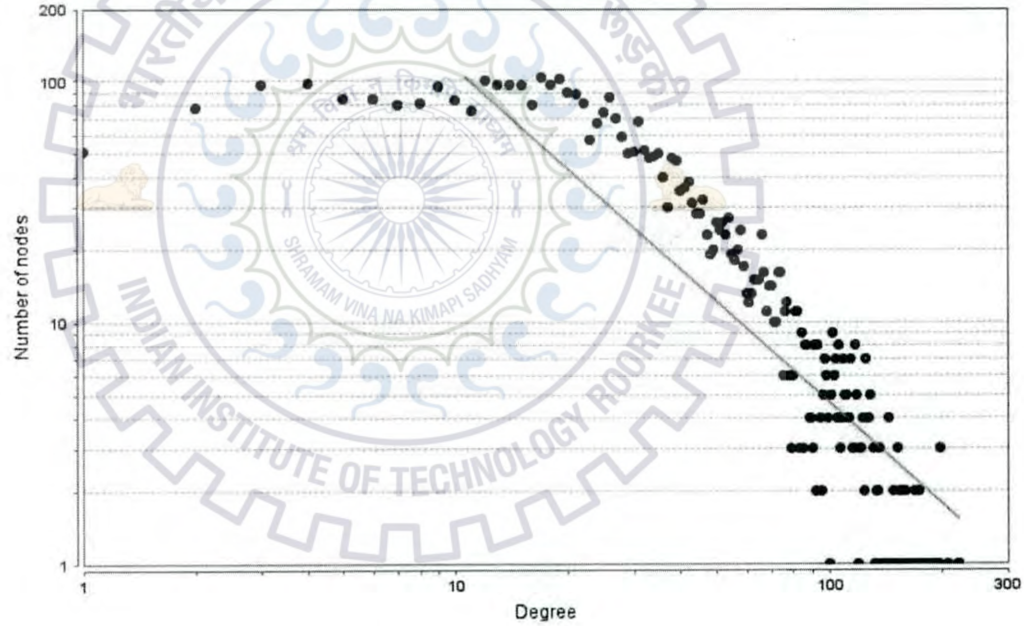


Figure 2.2 Node degree distribution. The distribution of the probability *p* (*k*) that the degree of a randomly chosen vertex equals *k* has been shown and it follows a power law.

As the degree distribution of the resulted network has also been shown in Figure 2.2, It exhibit scale-free property like many biological networks in which the degree distribution of proteins approximates a power law $P\ (k) = k^{-\gamma}$, with the degree exponent $\gamma \sim 1.38$. So that, there are very rare highly connected nodes called hubs in a vast majority of nodes with only a few connections. This means the network exhibits an important characteristic of being robust to random node failures [72]. On the other hand, scale-free networks are vulnerable to targeted attacks to hub nodes. Therefore, targeting central nodes would be important to alter the function of the network.

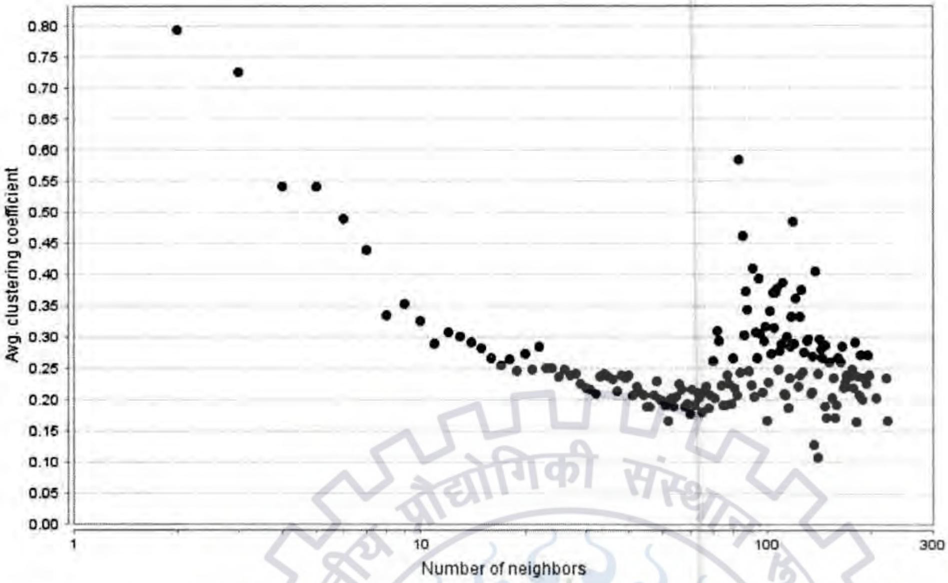


Figure 2.3 Average clustering coefficient distributions

Clustering coefficient is another basic statistical measure that accounts for the internal structure of a network [111]. It is related to the local cohesiveness of a network and measures the probability that two vertices with a common neighbour are connected. The clustering coefficient of the resulted network is significantly higher than the clustering coefficient of a random graph with the same number of vertices (0.008). Average clustering coefficient distribution of the network is shown in Figure 2.3.

2.3.2 Central proteins

The four centrality measures degree, betweenness, closeness and eigenvector have been used to rank proteins in the network. It is hypothesized that proteins which are central in the constructed disease-specific protein-protein interaction network are likely to be the potential drug targets of *Mycobacterium tuberculosis.* Since the resulted network is a scale free network it is vulnerable to targeted attacks on the central proteins. Which means it is possible to alter the function of the network by targeting these central proteins.

The focused is on the identification of proteins that found at the centre of gravity of the functional network which are with high betweenness and connected to some influential proteins at a certain levels. With this it is possible to identify proteins that are potentially essential for the survival of the bacterial pathogen, as they correspond to bottlenecks in the *Mycobacterium tuberculosis H37Rv* functional network and are, therefore expected to be key components of the organism's cellular processes. Bottleneck proteins are proteins responsible for several indirect functional connections between other proteins in the functional network [67]. Since the average shortest path length is 3.096, a protein in the functional network is said to belong to the gravity centre if its betweenness measure is above the total number of shortest paths expected to pass through the protein in the functional network of interest which is 12253.968. Based on this criterion, we got 807 ranked proteins of Mycobacterium Tuberculosis H37Rv.

2.3.3 Non-homologous proteins

The resulted central proteins were refined with non-human homologous protein dataset obtained from Drug Target Protein Database [97] since non-human homologous proteins are primary preference drug targets for host–parasite diseases like TB. Through this, a list of 390 non-human homologous proteins has been identified. These proteins were proposed as potential drug targets. It is believed that the list constitutes important proteins and thus, potential drug targets within the bacterial pathogen. The list of candidate proteins including their centrality values and overlaps with previously reported targets have been shown in Table 2.2.

Table 2.2 List of proposed targets. The table contains detailed list of proteins proposed as potential drug targets for *Mycobacterium tuberculosis H37Rv*. They have been sorted based on betweenness centrality measure. Structural information and validation, if these proteins were reported as drug targets have been incorporated.

Rv Number	Betweenness	Eigenvector	Degree	Closeness	Mazandu and Nulder	Uniprot Target List	TDR Posted	Drug Target Protein	Kings.et.al	3D Structure
Rv1303	115191.00	0.106641	207	0.049058		√				
Rv3527	98518.38	0.077701	182	0.048952						
Rv3705c	86579.07	0.093173	163	0.048807						
Rv3875	85161.14	0.020655	115	0.048651		√				√
Rv3231c	76353.84	0.029403	112	0.048591			√			
Rv2376c	75353.56	0.064592	154	0.048837						
Rv3474	73018.76	0.005009	61	0.048407			√			

Rv Number	Betweenness	Eigenvector	Degree	Closeness	Mazandu and Nulder	Uniprot Target List	TDR Posted	Drug Target Protein	Kings.et.al	3D Structure
Rv0910	70276.10	0.077285	141	0.048785						
Rv0311	67942.52	0.073912	156	0.048862			√	√		
Rv0451c	66230.34	0.096158	171	0.048834						
Rv0904c	64350.88	0.008161	94	0.048589		√				
Rv3905c	63660.71	0.043666	117	0.048657						
Rv0288	63475.45	0.091027	170	0.048855						
Rv1733c	62737.40	0.035406	117	0.048709						
Rv3288c	60230.99	0.104386	197	0.048911						
Rv1362c	60168.61	0.0672	157	0.048846						
Rv0096	59842.48	0.084631	144	0.048726				√		
Rv1808	58799.30	0.078432	152	0.048734						
Rv0506	58787.40	0.094458	178	0.048874						
Rv1599	58334.63	0.002786	82	0.04833			√			
Rv0058	57680.18	0.00275	62	0.048149		√		√		√
Rv2703	56665.16	0.00455	80	0.048385		√				
Rv0138	56539.35	0.042205	106	0.048705						
Rv3651	55481.38	0.06642	125	0.04867						
Rv2302	55353.41	0.014297	74	0.048432						√
Rv3492c	55311.20	0.073605	159	0.048821						
Rv1546	54574.60	0.057536	135	0.048855						
Rv0677c	52211.48	0.095557	169	0.048874						
Rv3287c	50641.43	0.065205	147	0.048881			√			
Rv1908c	50548.04	0.005638	76	0.048456		√			√	√
Rv2042c	50321.91	0.028753	93	0.048527						
Rv0549c	49954.80	0.004045	57	0.04794						
Rv1417	49386.11	0.038937	110	0.048601						
Rv0875c	49197.57	0.1194	197	0.048958			√			
Rv1271c	48622.22	0.052956	117	0.048634						
Rv0686	48556.00	0.069283	135	0.048773						

Rv Number	Betweenness	Eigenvector	Degree	Closeness	Mazandu and Nulder	Uniprot Target List	TDR Posted	Drug Target Protein	Kings.et.al	3D Structure
Rv1312	48275.48	0.054789	127	0.048801			√			
Rv1848	47954.98	0.003896	77	0.048443						√
Rv1274	46747.54	0.113241	197	0.049002			√			
Rv3493c	46593.20	0.105126	190	0.048919						
Rv1972	46383.98	0.026224	106	0.04853						
Rv3795	46319.86	0.053337	127	0.04885						
Rv3763	46305.10	0.062693	109	0.048592						
Rv1109c	45663.24	0.099705	176	0.048948						
Rv2137c	45030.71	0.067515	120	0.048758						
Rv0403c	43913.60	0.093394	156	0.048804						
Rv2144c	43811.76	0.091143	152	0.048821						
Rv0177	42918.04	0.091198	168	0.048892						
Rv1787	42807.19	0.057635	116	0.048609						
Rv3882c	42462.40	0.047485	113	0.048622		√				
Rv3355c	42306.69	0.060385	117	0.04873						
Rv1973	42228.27	0.03866	110	0.048569						
Rv2748c	41667.39	0.003447	54	0.048162	√	√				
Rv2468c	41600.56	0.090221	152	0.048792						
Rv2743c	41407.03	0.088029	151	0.048769						
Rv1390	41235.71	0.01245	128	0.048601			√			
Rv0817c	40821.99	0.119145	196	0.048974			√			
Rv1610	40741.26	0.059531	118	0.048729						
Rv1966	40339.45	0.010397	84	0.048499						
Rv0358	40230.57	0.079677	154	0.048828						
Rv2050	40088.31	0.019412	91	0.04851		√	√			
Rv1016c	39618.56	0.086168	142	0.048822						
Rv3849	39512.60	0.06444	112	0.048639		√				
Rv0979A	39316.26	0.016146	62	0.048332						
Rv1598c	39235.87	0.046498	114	0.048668						

Rv Number	Betweenness	Eigenvector	Degree	Closeness	Mazandu and Nulder	Uniprot Target List	TDR Posted	Drug Target Protein	Kings.et.al	3D Structure
Rv3457c	39017.70	0.015064	160	0.048611			√			
Rv2737c	38743.82	0.008274	113	0.048453						√
Rv2197c	38294.86	0.090015	136	0.048723						
Rv0941c	38289.07	0.032905	84	0.048458						
Rv2150c	37525.40	0.007633	103	0.048434			√			√
Rv3289c	37362.90	0.066157	136	0.048769						
Rv2330c	37196.41	0.049506	109	0.048632						
Rv2520c	37146.00	0.05662	124	0.048732			√			
Rv1332	37140.63	0.039947	101	0.048531			√			
Rv3874	37059.07	0.010753	75	0.048243						
Rv3793	36733.14	0.068721	136	0.04879		√				
Rv2074	36616.85	0.014234	77	0.048496						√
Rv0236A	36609.58	0.081771	131	0.048716						
Rv0010c	36431.21	0.072736	134	0.048816						
Rv3862c	36075.43	0.03167	104	0.048604						
Rv2111c	35411.24	0.004444	37	0.047943		√	√			
Rv3675	35375.73	0.051855	114	0.048642						
Rv3847	35141.36	0.050297	108	0.048624			√			
Rv2843	34959.63	0.037953	97	0.048545						
Rv3864	34797.55	0.011183	73	0.0482						
Rv0857	34491.38	0.033644	106	0.048636						
Rv1476	34379.21	0.092483	148	0.048805			√			
Rv2719c	34105.25	0.049744	99	0.048605						
Rv0262c	33648.72	0.024748	56	0.048232						√
Rv0002	33375.50	0.00362	90	0.04828						
Rv0885	32974.38	0.048192	91	0.048527			√			
Rv2239c	32695.59	0.012681	81	0.048472			√			
Rv3099c	32479.14	0.026729	84	0.048558						
Rv0258c	32203.14	0.024176	72	0.048393						

Rv Number	Betweenness	Eigenvector	Degree	Closeness	Mazandu and Nulder	Uniprot Target List	TDR Posted	Drug Target Protein	Kings.et.al	3D Structure
Rv0882	32105.29	0.058547	107	0.048555						
Rv3794	31967.25	0.057622	115	0.048761		√				
Rv2728c	31889.54	0.040382	74	0.048436			√			
Rv0883c	31779.75	0.050783	112	0.04874			√			
Rv3753c	31495.47	0.043638	102	0.048538						
Rv1980c	31354.69	0.007854	56	0.048217						√
Rv1591	31260.72	0.102258	165	0.048884						
Rv3219	31078.48	0.006471	69	0.048464			√			
Rv0179c	30756.63	0.04038	89	0.048471						
Rv0049	30554.35	0.041651	98	0.048604						
Rv1172c	30503.15	0.029209	63	0.048237						
Rv0760c	30486.24	0.017341	68	0.048345						√
Rv0513	29993.44	0.066879	121	0.04867						
Rv3892c	29968.37	0.015717	69	0.048358						
Rv0610c	29902.38	0.009353	41	0.048025						
Rv1891	29624.91	0.064194	105	0.04854						
Rv2077c	29608.41	0.031266	70	0.048321						
Rv3921c	29409.09	0.0068	68	0.048375						
Rv2206	28890.97	0.026674	82	0.048482						
Rv0079	28788.13	0.026357	76	0.048336						
Rv3820c	28730.83	0.019568	58	0.04818						
Rv3244c	28529.01	0.021865	64	0.048389			√			
Rv2462c	28445.73	0.01193	144	0.048393						
Rv1209	28409.20	0.056317	114	0.048569			√			
Rv1083	28283.58	0.050672	109	0.04861			√			
Rv1486c	27998.34	0.078581	125	0.048671			√			
Rv1571	27905.90	0.017814	59	0.048314						
Rv3597c	27758.72	0.018943	73	0.048406			√			
Rv3444c	27652.10	0.039839	101	0.048549			√			

Rv Number	Betweenness	Eigenvector	Degree	Closeness	Mazandu and Nulder	Uniprot Target List	TDR Posted	Drug Target Protein	Kings.et.al	3D Structure
Rv1558	27503.98	0.011858	67	0.048299		√				
Rv0569	27328.88	0.014829	74	0.048359						
Rv2256c	27194.08	0.018578	66	0.048292			√			
Rv2664	27192.28	0.018623	54	0.048129						
Rv1547	26895.52	0.00411	75	0.048315			√			
Rv0807	26830.41	0.017871	69	0.048465			√			
Rv1871c	26779.28	0.026794	76	0.048486						
Rv3707c	26771.50	0.048846	73	0.048371						
Rv3369	26766.04	0.013672	68	0.048358						
Rv0738	26763.20	0.011696	60	0.048289						
Rv0732	26715.15	0.011198	111	0.048285		√				
Rv2751	26543.23	1.06E-04	12	0.046424						
Rv3129	26530.56	0.018008	69	0.048247			√			
Rv0184	26432.16	0.051124	102	0.048597			√			
Rv2108	26251.40	0.082179	128	0.048704						
Rv0088	26082.72	0.023095	73	0.048288		√				
Rv3585	26068.18	0.003035	51	0.048052						
Rv0634B	26049.27	0.008919	87	0.048002						
Rv1875	26005.64	0.026246	81	0.048423			√			
Rv1727	25756.26	0.012359	52	0.048147						
Rv2659c	25629.14	0.001621	28	0.04776						
Rv2033c	25319.02	0.007829	60	0.04831						
Rv3662c	25068.74	0.013198	52	0.048151						
Rv2021c	24912.58	0.002246	38	0.04771						
Rv0185	24891.10	0.050268	102	0.048615						
Rv0200	24880.17	0.076966	127	0.048714						
Rv2043c	24872.61	0.004418	51	0.048254						
Rv3416	24851.81	0.01969	76	0.048463			√			
Rv2418c	24712.11	0.059897	116	0.048606			√			

Rv Number	Betweenness	Eigenvector	Degree	Closeness	Mazandu and Mulder	Uniprot Target List	TDR Posted	Drug Target Protein	Kings.et.al	3D Structure
Rv0453	24423.04	0.03976	64	0.048276				√		
Rv1837c	24302.38	0.002012	47	0.047937						√
Rv2186c	24297.74	0.086297	148	0.048809						
Rv3221A	24178.66	0.008386	47	0.048135						
Rv2695	24091.98	0.0715	128	0.048716			√			
Rv0178	24089.55	0.054647	103	0.048543						
Rv2778c	24030.64	0.022682	71	0.048359						
Rv0209	24024.48	0.037279	73	0.048467						
Rv3681c	23946.54	0.015426	76	0.048478			√			
Rv0467	23928.85	0.00294	53	0.048157						√
Rv1758	23804.38	0.008875	46	0.048076						
Rv3490	23777.69	7.30E-04	30	0.047695			√			
Rv1953	23685.74	1.49E-04	21	0.046261						
Rv1415	23617.26	0.00281	57	0.048159						
Rv3641c	23616.00	8.20E-05	4	0.045885						
Rv1356c	23586.40	0.03694	86	0.048497						
Rv0290	23576.32	0.073813	128	0.048616		√		√		
Rv2199c	23229.02	0.014123	74	0.048371						
Rv1959c	23205.33	1.77E-04	6	0.04653						
Rv0260c	23129.81	0.002362	48	0.048108						
Rv1780	23097.11	0.080416	126	0.048699						
Rv2219	23090.48	0.031297	101	0.04856						
Rv2762c	23056.36	0.007568	36	0.048042						
Rv2785c	23026.77	0.01109	125	0.048366						
Rv1650	22825.45	0.006996	104	0.048298		√				
Rv2906c	22799.92	0.011045	126	0.048353			√			
Rv3240c	22790.56	0.003691	64	0.048159			√		√	√
Rv0999	22690.63	0.079325	128	0.048687						
Rv2534c	22585.90	0.010743	123	0.048378			√			

Rv Number	Betweenness	Eigenvector	Degree	Closeness	Mazandu and Nulder	Uniprot Target List	TDR Posted	Drug Target Protein	Kings.et.al	3D Structure
Rv1203c	22510.49	0.045064	99	0.048541						
Rv2844	22433.52	0.029142	62	0.048375						√
Rv1444c	22397.98	0.04098	89	0.048483						
Rv0227c	22294.80	0.075547	111	0.048643				√		
Rv0909	22265.93	0.025797	60	0.048257						
Rv2598	22240.34	0.014127	57	0.048221						
Rv3718c	22127.45	0.030159	84	0.048519			√			
Rv3780	22001.60	0.016156	70	0.048352			√			
Rv3201c	21984.81	0.006371	39	0.048036			√			
Rv1382	21967.10	0.032071	87	0.048535						
Rv2868c	21919.65	0.002969	61	0.048131			√			
Rv1738	21804.76	0.01357	66	0.048355		√				
Rv1508c	21686.20	0.005711	50	0.048158						
Rv0707	21619.61	0.012703	138	0.048321						
Rv3773c	21575.53	0.014991	66	0.048333						
Rv0613c	21478.25	0.009386	41	0.048078						
Rv2093c	21413.89	0.002743	52	0.048147						
Rv3463	21357.19	0.011208	50	0.048205						
Rv0493c	21213.66	0.014184	57	0.04819						
Rv2553c	21154.97	0.004896	53	0.048296						
Rv0048c	20952.14	0.070939	118	0.048642						
Rv2203	20893.83	0.045379	67	0.048359						
Rv1118c	20885.30	0.022408	66	0.048304						
Rv1363c	20883.88	0.041346	97	0.048589						
Rv1368	20876.93	0.018631	50	0.048263						
Rv3616c	20855.97	0.017509	69	0.048237						
Rv2772c	20700.48	0.028871	69	0.048398						
Rv3658c	20682.94	0.043273	87	0.04855						
Rv0216	20616.27	0.019399	57	0.048213			√			√

Rv Number	Betweenness	Eigenvector	Degree	Closeness	Mazandu and Mulder	Uniprot Target List	TDR Posted	Drug Target Protein	Kings.et.al	3D Structure
Rv3839	20518.78	0.031363	67	0.048321						
Rv0431	20463.85	0.020625	42	0.048121						
Rv1797	20461.07	0.060734	102	0.048522			√			
Rv1602	20447.48	0.002364	49	0.048106						
Rv2945c	20284.02	0.021909	61	0.048257					√	√
Rv0365c	20274.93	0.022385	54	0.048173						
Rv0481c	20220.11	0.052012	110	0.04859						
Rv3205c	20169.22	0.036587	82	0.048417			√			
Rv1794	19814.71	0.050454	88	0.048461						
Rv3413c	19691.26	0.023709	65	0.048336			√			
Rv2700	19611.58	0.080525	132	0.048724						
Rv2638	19572.07	0.012337	56	0.048163						
Rv2097c	19334.73	0.007278	57	0.048266			√			
Rv3412	19248.75	0.046827	102	0.048545			√			
Rv2647	19166.78	0.001202	12	0.047157						
Rv3802c	19149.55	0.026657	65	0.048293			√			
Rv1722	19135.50	0.001015	15	0.047519						
Rv3912	19116.66	0.031938	75	0.048397						
Rv1044	19090.46	0.003662	25	0.047656						
Rv3887c	19033.49	0.015185	63	0.048228						
Rv3133c	18975.65	0.003374	50	0.048136		√				√
Rv1148c	18826.16	0.033097	78	0.04838						
Rv2558	18800.55	0.030753	84	0.04842						
Rv0292	18595.61	0.057754	93	0.048449				√		
Rv0280	18529.92	0.026589	57	0.048134		√		√		
Rv2942	18515.60	0.005235	46	0.047829						
Rv3143	18413.58	0.001627	36	0.047833						
Rv3180c	18391.19	0.005237	36	0.047854						
Rv3877	18375.84	0.015738	63	0.04823						

Rv Number	Betweenness	Eigenvector	Degree	Closeness	Mazandu and Nulder	Uniprot Target List	TDR Posted	Drug Target Protein	Kings.et.al	3D Structure
Rv3395c	18336.81	0.030765	78	0.04838						
Rv0531	18332.45	0.021028	50	0.048157						
Rv3202c	18322.98	0.005936	38	0.047981						
Rv2813	18302.37	0.001109	28	0.047721						
Rv3439c	18150.66	0.03165	77	0.048395						
Rv3279c	18050.24	0.002904	46	0.048135						√
Rv1159	17943.38	0.00711	41	0.047953						
Rv2880c	17896.85	0.003893	69	0.048104						
Rv0487	17873.10	0.020242	65	0.048342			√			
Rv2492	17856.27	0.002282	27	0.047683						
Rv2552c	17841.86	0.002641	61	0.048196						
Rv2655c	17806.45	0.003903	35	0.04778						
Rv1611	17798.76	0.002891	57	0.048196		√	√			
Rv3065	17756.42	7.77E-04	33	0.047492						
Rv1706c	17578.46	0.036745	79	0.048333		√				
Rv1691	17516.41	0.007204	52	0.04825						
Rv2847c	17485.74	0.00171	56	0.047952						
Rv0313	17446.45	0.041132	88	0.048563						
Rv2570	17407.03	0.005074	24	0.047774						
Rv0490	17383.59	0.001862	45	0.047885						
Rv2773c	17351.88	0.001899	52	0.048026						√
Rv3584	17278.03	0.038645	80	0.04843						
Rv0249c	17255.22	0.020176	61	0.048314						
Rv0093c	17217.30	0.039443	69	0.048297						
Rv0080	17182.14	0.012088	54	0.048138						
Rv3792	17136.64	0.036626	76	0.048349						
Rv2474c	17136.04	0.033316	70	0.04841						
Rv1315	16958.69	0.003925	66	0.048088			√			
Rv0955	16952.05	0.046121	72	0.048436						

Rv Number	Betweenness	Eigenvector	Degree	Closeness	Mazandu and Nulder	Uniprot Target List	TDR Posted	Drug Target Protein	Kings.et.al	3D Structure
Rv1258c	16939.99	0.001387	34	0.047678						
Rv3321c	16821.23	0.001082	27	0.047456						
Rv2421c	16735.58	0.001805	49	0.047984			√			
Rv0863	16658.38	0.026374	67	0.048345			√			
Rv0902c	16640.24	0.021116	42	0.048067						√
Rv2848c	16638.95	0.001328	50	0.047756			√			
Rv1402	16635.79	0.004491	67	0.04818						
Rv1711	16528.08	0.001896	59	0.048038			√			
Rv0361	16421.20	0.070502	102	0.048581						
Rv3363c	16373.15	0.03274	74	0.048389						
Rv1696	16237.53	0.003924	72	0.048178						
Rv1694	16224.56	0.002499	58	0.048124		√	√			
Rv1343c	16204.04	0.048565	104	0.04862						
Rv0289	16198.58	0.075002	108	0.04855			√	√		
Rv1024	16154.03	0.009734	52	0.048235			√			
Rv2632c	16130.89	0.012548	49	0.048325						√
Rv1115	16069.72	0.005738	27	0.04772			√			
Rv1252c	15874.61	0.023335	49	0.048194						
Rv3647c	15860.47	0.055002	89	0.048554						
Rv2599	15859.13	0.019286	58	0.048228						
Rv2629	15853.57	0.010869	52	0.048161						
Rv3723	15825.21	0.058028	84	0.048402						
Rv2169c	15765.12	0.016853	49	0.048135						
Rv3483c	15759.64	0.016006	54	0.048288						
Rv0245	15748.00	1.15E-05	2	0.044801			√			
Rv3767c	15748.00	1.21E-06	2	0.044376						
Rv2270	15698.57	0.001401	17	0.047193						
Rv2711	15661.20	0.005903	41	0.048172			√			√
Rv0651	15638.88	0.012552	118	0.048362			√			

Rv Number	Betweenness	Eigenvector	Degree	Closeness	Mazandu and Nulder	Uniprot Target List	TDR Posted	Drug Target Protein	Kings.et.al	3D Structure
Rv3843c	15604.25	0.049049	76	0.048473						
Rv0474	15407.40	0.001318	20	0.047424						
Rv2311	15368.34	0.001095	10	0.046828						
Rv3601c	15341.17	0.002516	49	0.048123						
Rv0495c	15270.48	0.030429	76	0.04845						
Rv2112c	15258.44	0.008602	58	0.0483		√	√			
Rv2191	15224.33	0.008908	45	0.048355						
Rv3252c	15193.15	0.001787	37	0.047804						
Rv1280c	15176.18	0.013402	39	0.047987						
Rv0383c	15151.17	0.026303	65	0.048319				√		
Rv2009	15111.71	0.001923	31	0.047396						
Rv3437	15044.15	0.021755	70	0.048339						
Rv1265	15017.98	0.014732	52	0.048186						
Rv3223c	15007.41	0.001591	38	0.047918		√	√			
Rv2518c	14995.93	0.002819	25	0.047732			√			
Rv0745	14858.4	0.006472	30	0.047828			√			
Rv3531c	14827.18	0.009694	42	0.048042						
Rv0233	14782.64	0.001912	29	0.047897					√	√
Rv2412	14768.68	0.007477	73	0.047919		√	√			
Rv2072c	14766.83	0.0015	48	0.047811						
Rv0295c	14745.90	0.013497	34	0.04799						
Rv0533c	14668.97	0.002576	53	0.048013		√			√	√
Rv3660c	14637.48	0.009687	43	0.048086		√	√			
Rv2721c	14595.70	0.009142	30	0.047907						
Rv0679c	14591.15	0.007703	27	0.047816						
Rv0713	14530.87	0.028288	61	0.048236						
Rv2620c	14514.36	0.013896	53	0.048114						
Rv3635	14474.55	0.049931	77	0.048468						
Rv2484c	14456.00	0.013058	44	0.048024						

Rv Number	Betweenness	Eigenvector	Degree	Closeness	Mazandu and Nulder	Uniprot Target List	TDR Posted	Drug Target Protein	Kings.et.al	3D Structure
Rv2525c	14428.34	0.042225	82	0.04847			√			
Rv1302	14417.00	0.005331	56	0.048101						
Rv3547	14295.91	0.007095	50	0.048195						
Rv2604c	14265.08	0.004825	43	0.048096						
Rv0736	14264.87	0.015363	50	0.048301			√			
Rv0065	14263.71	0.003886	30	0.047682						
Rv2939	14256.58	0.00668	41	0.047849		√				√
Rv3850	14249.49	0.059728	98	0.048556			√			
Rv3809c	14229.11	0.004649	43	0.047792	√		√			√
Rv2538c	14206.77	0.002325	65	0.048107		√				
Rv1970	14202.21	0.008613	60	0.048067						
Rv0164	14149.58	0.020211	39	0.048144						
Rv1845c	14084.02	0.018668	44	0.048172						
Rv1860	14040.37	0.008405	44	0.047985						
Rv1387	13983.82	0.046963	74	0.048399						
Rv3482c	13941.59	0.010446	65	0.048239						
Rv1294	13762.81	0.001326	51	0.047978			√			
Rv0670	13708.37	0.001171	46	0.047865						
Rv2151c	13691.68	0.008896	54	0.048147			√			
Rv0040c	13573.57	0.018812	50	0.048173						
Rv1823	13558.92	0.010927	44	0.048024						
Rv0408	13516.32	0.002005	58	0.048091						
Rv2185c	13452.28	0.012208	46	0.048171			√			
Rv2416c	13419.65	0.007718	39	0.048289						
Rv3885c	13385.35	0.017462	52	0.048162						
Rv0963c	13347.38	0.010121	44	0.048092						
Rv1633	13298.87	0.002318	59	0.048006						
Rv3100c	13271.99	0.005532	74	0.048093						
Rv0066c	13193.77	8.49E-04	34	0.047694						

Rv Number	Betweenness	Eigenvector	Degree	Closeness	Mazandu and Nulder	Uniprot Target List	TDR Posted	Drug Target Protein	Kings.et.al	3D Structure
Rv0934	13192.23	0.002668	37	0.047857						√
Rv0561c	13191.79	0.003724	36	0.047852			√			
Rv0621	13184.42	0.006122	21	0.047755						
Rv2987c	13133.91	0.001755	52	0.047983						
Rv2986c	13109.86	0.003152	40	0.048122		√				√
Rv0344c	13059.52	0.047784	82	0.048423						
Rv0419	12947.32	0.025772	49	0.048153						
Rv3033	12904.42	0.019676	57	0.048358						
Rv2667	12860.25	0.001834	30	0.047825						
Rv2068c	12845.59	0.004436	31	0.048003						√
Rv1712	12815.69	0.002192	62	0.04802						
Rv0286	12808.90	0.039534	75	0.048263		√		√		
Rv2231c	12805.97	0.002218	39	0.047811						
Rv0321	12756.10	0.001646	33	0.047857			√			√
Rv1365c	12714.26	0.014089	47	0.048178						
Rv0475	12603.79	0.007643	35	0.048034						
Rv0949	12572.48	0.00179	46	0.047999						
Rv2301	12541.64	0.014711	40	0.048212						
Rv0867c	12540.09	0.012514	57	0.048167						
Rv1632c	12535.47	0.03668	70	0.04841						
Rv3434c	12524.90	0.017269	55	0.048288						
Rv0442c	12489.61	0.002528	31	0.047616				√		
Rv0054	12459.25	0.002642	49	0.048052						√
Rv0256c	12440.24	0.020005	42	0.048102		√		√		
Rv0026	12424.85	0.004571	31	0.047654						
Rv2391	12401.70	0.001206	45	0.047875		√	√			√
Rv3258c	12374.61	0.012237	60	0.048281			√			
Rv1783	12333.33	0.010263	46	0.048001						
Rv2610c	12327.83	0.004189	45	0.047918						

Rv Number	Betweenness	Eigenvector	Degree	Closeness	Mazandu and Nulder	Uniprot Target List	TDR Posted	Drug Target Protein	Kings.et.al	3D Structure
Rv1940	12319.11	0.002024	40	0.047992						
Rv2712c	12306.36	0.026541	70	0.048351						
Rv0133	12304.99	0.001614	45	0.047579						
Rv0389	12291.33	0.001289	33	0.04774						
Rv0954	12258.49	0.013444	40	0.048003						
Rv2693c	12257.61	0.021034	63	0.048328						

The computational validation of identified set of potential drug targets is difficult and there are also other factors like the availability and predictability of three-dimensional structures of protein targets which affects the suitability of drug targets in addition to the targets having a potentially important functional role [67]. However, we tried to assess the candidate list by searching from literatures and by using *mycobacterium tuberculosis* drug target databases. We have used TDR posted drug targets of *mycobacterium tuberculosis* [104] which contains validated drug targets obtained though manual curation from literatures. We also identified a high confidence drug targets from UniProt annotation for *Mycobacterium tuberculosis H37Rv* [103] and looked at a handful of genes reported to be predicted drug targets by Kinnings et al. [112]. Additionally, we have used the top 10 candidates from Mazandu and Nulder [67]. As it has been shown in Figure 2.4 our lists of potential drug targets includes 85 proteins which were TDR validated targets, 9 of which were in UniProt target list, 2 of which in Drug Target Protein Database [97] and 1 of each in Mazandu and Nulder, and Kinnings et al. An additional of 26 proteins from our lists were overlapped with UniProt target list, 5 of which were in Drug Target Protein Database, 2 of which in Kinnings et al. and 1 of which in Mazandu and Nulder. Our list also includes additional 6 proteins from Drug Target Protein Database and 2 proteins from Kinnings et al. Therefore out of our lists of 390 non-human homologous potential protein targets, 119 (30.51%) have previously been predicted or reported to be drug targets. Alternatively, the lists of these overlapping proteins can be referred from Table 2.2.

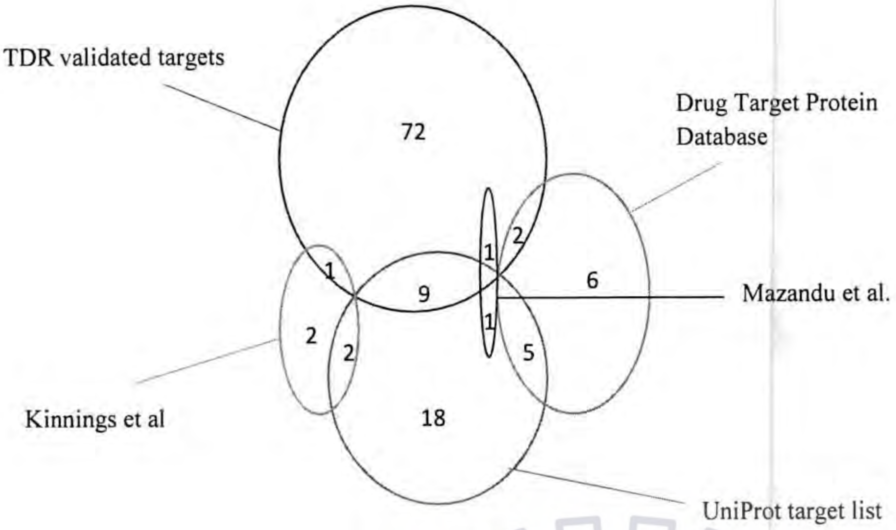


Figure 2.4 Venn diagram of candidates in our list which were reported by other methods

Table 2.3 Non-reported proteins in the top 10% of all of the four centrality measures

Uniprot acc	Gene name	Functional class	Network centrality Scores			
			Betweenness	Eigenvector	Degree	Closenes
P71874	Rv3527	Conserved Hypotheticals	98518.38	0.077701	182	0.048952
O69673	Rv3705c	Conserved Hypotheticals	86579.07	0.093173	163	0.048807
P9WJ05	Rv0910	Conserved Hypotheticals	70276.10	0.077285	141	0.048785
P9WJS9	mmpS4	Cellwall and Cell Processes	66230.34	0.096158	171	0.048834
P9WNK3	esxH	Cellwall and Cell Processes	63475.45	0.091027	170	0.048855
Q7D5S0	usfY	Conserved Hypotheticals	60230.99	0.104386	197	0.048911
P9WJT3	mmpS2	Cellwall and Cell Processes	58787.40	0.094458	178	0.048874
P9WJS7	mmpS5	Cellwall and Cell Processes	52211.48	0.095557	169	0.048874
O06356	Rv3493c	Cellwall and Cell Processes	46593.20	0.105126	190	0.048919

We have also closely observed the identified proteins which are not reported by the mentioned methods as drug target candidate for *Mycobacterium tuberculosis* and found on the top 27 (10%) in all of the four centrality measures that we have used. Based on this criterion, we have got 9 proteins that need further detail computational or experimental investigations. These proteins have been shown in Table 2.3.

2.3.4 Structural coverage of proposed targets

The availability and predictability of three dimensional structures for the proposed targets is one of the major factors which affect the druggablity. From 3,999 proteins in the *Mycobacterium tuberculosis* proteome, only 324 unique proteins have solved structure in the RCSB Protein Data Bank (PDB) [92]. This is approximated to only 8.1% structural coverage. However, by taking reliable homology models into consideration, it is possible to increase the structural coverage of the *Mycobacterium tuberculosis* proteome [112]. Out of 390 proteins from our proposed target list, only 33 have solved structure which is approximated to 8.46%.

2.4. Conclusion

In this study, a list of non-homologous protein targets for *Mycobacterium tuberculosis H37Rv* has been identified based on protein-protein interaction network analysis. The protein-protein interaction network has been constructed using a dataset obtained from STRING database [96]. A dataset containing list of non-homologous proteins of *mycobacterium tuberculosis*, retrieved from Drug Targets Protein Database [97], has been used because protein that are non-homologous to the host are primary preference drug targets in host–parasite diseases like TB. Next, we used degree, eigenvector, closeness and betweenness centrality metrics to rank the proteins in the network according to their relevance to the disease. We hypothesized that proteins that are central in the constructed disease specific network are the ones contributing to the survival of the bacterial pathogen within the host and they are likely to be potential drug targets.

We tried to validate the predicted candidate list by using already identified drug targets with the help of drug target databases of mycobacterium tuberculosis and literature reports. Some of the known targets for existing anti-tubercular drugs are not in the essential target lists, and some of the previously reported targets are not necessarily the most central or influential. This could be due to the focus on non- homologous protein lists in this investigation or there does not appear to be a single rule for identifying the best targets. However, this may contribute to the process of developing new antibiotics with novel mechanisms of action for better treatment of the disease by saving time and reducing cost.

CHAPTER THREE

Comparative Genome and Network Centrality Analysis to Identify Drug Targets of *Mycobacterium Tuberculosis H37Rv*

3.1 Introduction

Computational methods usually identify a list containing larger number of potential drug targets. Validating these targets with the aid of experiments could be very difficult task mainly due to time and cost constraints. Which means various ways and mechanisms should be implemented to validate and minimize the list of drug targets. In the previous analysis, most central proteins were identified from protein-protein interaction network of *Mycobacterium tuberculosis H37Rv* and those which are non homologous to human were proposed as a list of potential drug targets. As an extension to this analysis, we use an integration of essentiality dataset, non homology assessment and global network centrality measures in the current work by hypothesising that this integration would give us a list of more reliable potential drug targets. Identifying genes which are essential for the survival and growth of the organism and proposing them as potential drug targets is one of the popular approaches in drug target discovery especially for host–parasite diseases like tuberculosis.

There are three main findings which proposed the lists of essential genes for the survival and growth of *Mycobacterium tuberculosis H37Rv* [113-115]. Sassetti et al. (2003) identified a comprehensive list of genes that are essential for the optimal growth of MTB through transposon site hybridization (TraSH) [113]. They have found out that most of the essential genes for MTB growth were conserved in the degenerate genome of the leprosy bacillus and *Mycobacterium leprae.* This means there were losses of non-essential functions due to bacterium diverged from other mycobacteria. On the contrary, many of the identified genes don't have identifiable orthologues in other bacteria, indicating that there is a variation among organisms with different evolutionary histories in the requirement of minimum set of genes for survival. In the analysis by Griffin et al. (2011), global phenotypic profiling has been used to identify genes required for the growth of *Mycobacterium tuberculosis* where the composition of complex mutant libraries was characterized by using a combination of high-density mutagenesis and deep-sequencing [114]. The method is stated to be significant advancement over the prior methods employed for the identification of essential genes. Zhang et al. (2012) identified a comprehensive list of essential genes for the optimal growth of *Mycobacterium tuberculosis* through an unbiased high density mutagenesis and deep sequencing analysis [115].

The stated lists of essential genes from the three findings were compiled and stored in Database of Essential Genes (DEG) for the intended users [98-100]. DEG has been used to propose potential drug targets of *Mycobacterium tuberculosis H37Rv* [30]. In this study, the complete genome of *Mycobacterium tuberculosis H37Rv* was blasted against DEG to identify essential genes and the resulted dataset was further analyzed for similarity search against human genome to identify genes which are not similar with human to avoid host toxicity. Since two of the main findings about the essential genes were published after this study, it is possible to hypothesis that a more comprehensive set of potential drug targets of *Mycobacterium tuberculosis H37Rv* could be obtained through a systematic computational analysis on the integrated dataset from DEG which incorporates those recent findings.

From the analysis, a list of 137 potential drug targets of *Mycobacterium tuberculosis H37Rv* has been identified. These proteins are essential for the growth and survival of the pathogen, non homologous with human and prioritized based on their network centrality measure values where all of them found within the close neighbourhood of the centre of gravity of protein-protein interaction network. It has been found out that almost half of these proteins have been already reported as potential drug targets of the pathogen by other methods. The structural assessment showed that 28 out of the 137 (20.44 %.) proteins have solved structure.

3.2 Materials and Method

The complete genome sequence dataset of *Mycobacterium tuberculosis H37Rv* was retrieved from Tuberculosis Database which is an integrated platform providing access to genome sequence, expression data and literature curation for tuberculosis researches [28, 95]. BLAST search of the retrieved protein coding genes was carried out against DEG to identify essential genes. The corresponding protein sequences obtained after DEG search were subjected to a BLASTp against the non-redundant database with an e-value threshold cut off set to 0.005 [88]. The search was also restricted to H. sapiens because the objective was to find only those proteins, which do not have detectable human homologues to prevent host toxicity.

Subsequently, the protein-protein interaction network of *Mycobacterium tuberculosis H37Rv* was retrieved from Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) database [96]. Obviously, protein-protein interaction network datasets of this type contains false positives and false negatives which might affect the quality of the dataset and have an impact on the result of analysis based on it. To minimize this impact, interactions labeled with only "medium confidence" and "high confidence "scores were considered. The resulted lists of proteins were then prioritized by using the four network centrality measures such as degree, closeness, betweenness and eigenvector. The progression of the analysis in this study has been shown in Figure 3.1.

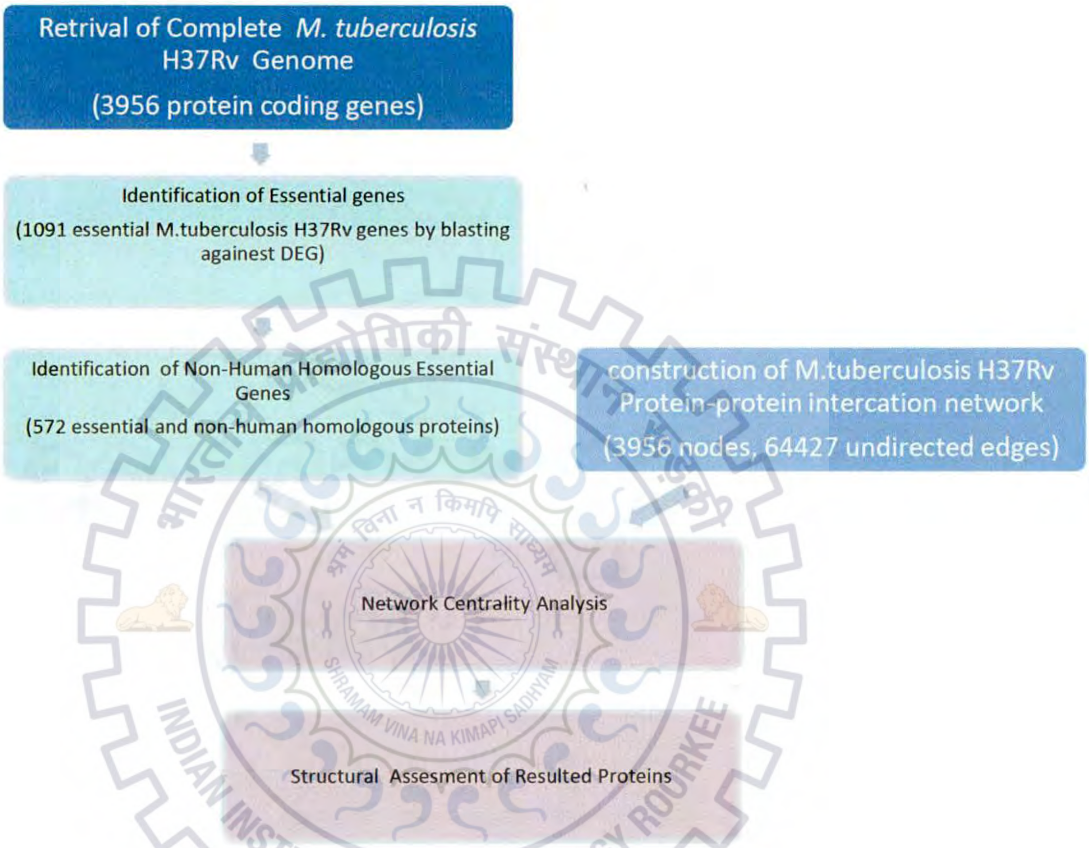

Retrival of Complete *M. tuberculosis* H37Rv Genome
(3956 protein coding genes)
Identification of Essential genes
(1091 essential M.tuberculosis H37Rv genes by blasting againest DEG)
Identification of Non-Human Homologous Essential Genes
(572 essential and non-human homologous proteins)
construction of M.tuberculosis H37Rv Protein-protein intercation network
(3956 nodes, 64427 undirected edges)
Network Centrality Analysis
Structural Assesment of Resulted Proteins

Figure 3.1 Progression of experiments. Different aspects indicated in this diagram are identification of essential genes, comparative analysis, construction of protein-protein interaction network, network centrality analysis and validation.

3.3 Results and Discussion

3.3.1 Comparative analysis for identifying non-homologous essential genes

The retrieved complete genome sequence dataset of *Mycobacterium tuberculosis H37Rv* consists of sequences of 3956 protein coding genes. These genes were then blasted against DEG to obtain essential genes. These genes are indispensable for the survival and growth of the pathogen. As a result their functions are, therefore, considered a foundation of life. Defining the protein coding genes which are essential for the bacterial growth and its survival is believed to be important in identifying both key biological processes and potential targets for rational drug development [115]. A total of 1091 genes were identifies as an essential genes from the analysis.

One of the important question that needs to be addressed while choosing proteins as potential drug targets for pathogens like *Mycobacterium tuberculosis H37Rv* is validating whether the proteins to be targeted are all absent in the host, *H. sapiens* and therefore unique to the pathogen. Identifying enzymes of the pathogen that are non-homologous with the host proteins as potential drug targets ensures that the targets have nothing in common with the host proteins, thereby, eliminating undesired host protein–drug interactions. We have performed a comparative analysis of the host, *H. sapiens* and the pathogen, *mycobacterium tuberculosis* for the identified 1091 essential genes. We have adopted a stringent measure of listing out only those proteins which have no similarity or negligible similarity (above the *e*-value threshold of 0.005) to the host proteins. With the aid of this approach, 572 out of 1091 proteins are absent in host *H. sapiens.* Therefore, these proteins are unique to *Mycobacterium tuberculosis H37Rv.*

3.3.2 Interactome network

A proteome-scale interaction network of *Mycobacterium tuberculosis H37Rv* was constructed using a dataset retrieved from STRING database [96]. The existence of false positives and false negatives is widely anticipated in a network of this type [44]. It has also been indicated that the protein-protein interaction networks generated from STRING database are of low quality consisting of a significant amount of false positives and false negatives [110]. Thus, if careful consideration has not been taken in using interactions derived from this dataset, there would be an impact on the result of the analyses. All interactions with value of 'low-scores' have been removed from the study to minimize the impact of the problem. The resulting network contains 64,427 interactions among 3,956 proteins. Of the total 64,427 interactions, 22,528 were labeled as 'high-score' and 41,899 as 'medium-score'. Despite of its shortcomings, this network provides a good framework for navigation through the proteome and it also allows for refinement of the network upon the availability of new experimental data.

3.3.3 Network Centrality Analysis

Comparative genome analysis was helpful in filtering out 572 non-homologous and essential proteins of *Mycobacterium tuberculosis H37Rv.* However, this set is still very large to validate with the aid of experimental methods. The network centrality measures have been used for further ranking and prioritizing these proteins in the generated proteome interactome network. The objective was to sort the resulted non-homologous and essential proteins based on their network centrality values so that highly ranked proteins can be used first in the development of drugs for the pathogen. This has been done with the subsequent steps of ranking all of the proteins in the generated network, filtering proteins that found near to the center of gravity and identifying the ordered list from non-homologous and essential proteins found in the filtered list. For longitudinal comparison of centralities, the distribution of betweenness value of sorted proteins has been indicated in Figure 3.2. The diagram shows the number of proteins located in separate score intervals of the network. Betweenness centrality metric is one of a significant indicator of network essentiality because proteins with high betweenness are essential for the functioning of the system by serving as a bridge of communication between several other proteins in the network [72].

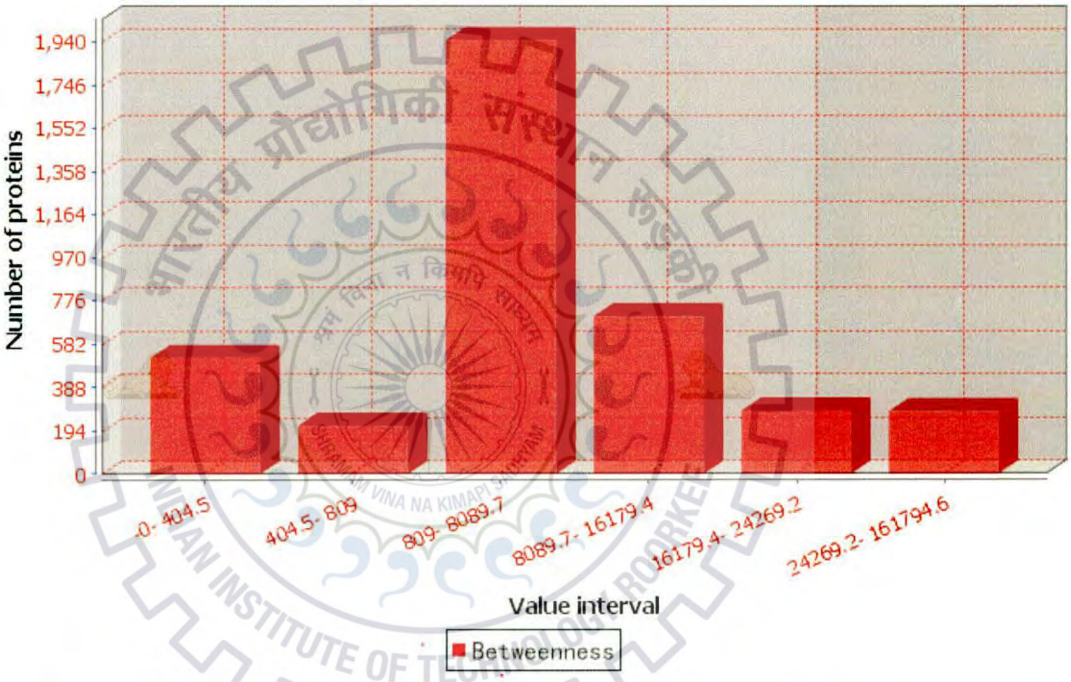


Figure 3.2 Distribution of betweenness centrality values. Longitudinal comparison of centrality values which shows the number of proteins located in separate score intervals based on betweenness centrality measure.

In this investigation, we tried to identify proteins which are found near to the centre of gravity of the proteome network by being connected with influential proteins. Since the characteristic path length of the generated network is 3.096, a protein is said to be at the centre of gravity if its betweenness measure is above the total number of shortest paths expected to pass through the protein in the functional network of interest, which is 12253.968. This criterion has been used by Mazandu and Nulder (2012) in identification of potential drug targets of *Mycobacterium tuberculosis* [67]. With the aid of this principle we have got 137 ranked, essential, non-homologous and central proteins which we believe are reliable targets for *Mycobacterium tuberculosis H37Rv*. The detail list of these potential drug targets incorporating the network centrality measure scores and their function is provided in Table 3.1. The lists of refined targets were further categorized by high level of functional classes and the distribution of these potential drug targets per functional class is shown in Figure 3.3. The distribution indicated that most of the candidate drug targets are involved in cell wall and cell processes, followed by a significant proportion of proteins in intermediary metabolism and respiration, conserved hypothetical and those belong to information pathway.

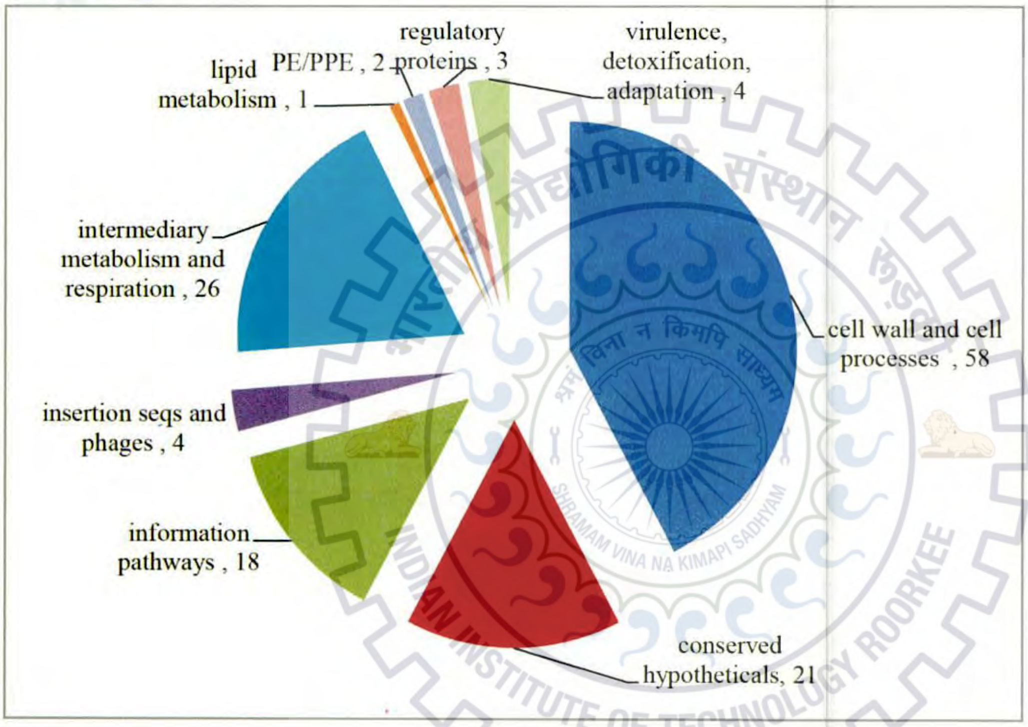


Figure 3.3 High level functional class distributions of proposed drug targets. The candidate lists have been classified into their high level functional class and the distribution has been indicated in the diagram.

Table 3.1 Proposed targets. Prioritized and detailed list of proteins proposed as potential drug targets for *Mycobacterium tuberculosis H37Rv* from this analysis.

Rv Number	Function	Betweenness	Eigenvector	Degree	Closeness
Rv1303	unknown	115190.99	0.10664081	207	0.04905838
Rv3019c	unknown	71224.330	0.08889492	167	0.0488181
Rv0311	unknown	67942.516	0.07391167	156	0.048861504
Rv0556	unknown	66400.050	0.103606604	193	0.049006738
Rv0451c	unknown	66230.336	0.09615834	171	0.04883437
Rv0288	May be involved in virulence.	63475.453	0.09102694	170	0.048855472
Rv1599	Involved in histidine biosynthesis pathway (tenth step). This protein is considered as a bifunctional enzyme, possessing two active sites, one an alcohol dehydrogenase and the other an aldehyde dehydrogenase [catalytic activity: L-histidinol + 2 NAD(+) + H(2)O = L-histidine + 2 NADH].	58334.630	0.002786182	82	0.048330363
Rv0058	Participates in initiation and elongation during chromosome replication; it exhibits DNA-dependent ATPase activity. The intein is an endonuclease (potential).	57680.180	0.002750394	62	0.048148643
Rv2703	The sigma factor is an initiation factor that promotes attachment of the RNA polymerase to specific initiation sites and then is released. This is the primary sigma-factor of this bacteria. Supposedly involved in the housekeeping regulons.	56665.156	0.004549677	80	0.04838473
Rv3859c	Probably involved in glutamate biosynthesis [catalytic activity: 2 L-glutamate + NADP(+) = L-glutamine + 2-oxoglutarate + NADPH].	54468.613	0.002801649	93	0.048351623
Rv1908c	Multifunctional enzyme, exhibiting both a catalase, a broad-spectrum peroxidase, and a peroxynitritase activities. May play a role in the intracellular survival of mycobacteria within macrophages; protection against reactive oxygen and nitrogen intermediates produced by phagocytic cells. Seems regulated by SIGB|Rv2710 [catalytic activity: 2 H(2)O(2) = O(2) + 2 H(2)O].	50548.035	0.005637622	76	0.048456423
Rv0875c	unknown	49197.566	0.11939961	197	0.04895823
Rv1274	unknown	46747.540	0.11324064	197	0.04900249
Rv3795	Involved in the biosynthesis of the mycobacterial cell wall arabinan and resistance to ethambutol (EMB; dextro-2,2'-(ethylenediimino)-DI-1-butanol). Polymerizes arabinose into the arabinan of arabiogalactan [catalytic activity: UDP-L-arabinose + indol-3-ylacetyl-myo-inositol = UDP + indol-3-ylacetyl-myo-inositol L-arabinoside].	46319.860	0.0533371	127	0.04885004

Rv Number	Function	Betweenness	Eigenvector	Degree	Closeness
Rv3604c	unknown	43248.620	0.045404814	99	0.048615973
Rv2748c	Possibly involved in cell division processes	41667.390	0.003446955	54	0.048161536
Rv1390	Promotes RNA polymerase assembly. Latches the N-and C-terminal regions of the beta' subunit thereby faciltating its interaction with the beta and alpha subunits [catalytic activity: N nucleoside triphosphate = N diphosphate + {RNA}N].	41235.707	0.012449709	128	0.048601046
Rv0817c	unknown	40821.99	0.11914455	196	0.048973985
Rv1610	unknown	40741.258	0.059531	118	0.048728526
Rv0358	unknown	40230.57	0.07967692	154	0.048827738
Rv2050	unknown	40088.305	0.019411962	91	0.04851048
Rv3457c	DNA-dependent RNA polymerase catalyzes the transcription of DNA into RNA using the four ribonucleoside triphosphates as substrates. The amino-terminal portion is involved in the assembly of core RNAP, whereas the C-terminal is involved in interaction with transcriptional regulators [catalytic activity: N nucleoside triphosphate = N pyrophosphate + RNA(N)].	39017.700	0.015064457	160	0.048611198
Rv2150c	Essential for cell division. It is thought that the intracellular concentration of FTSZ protein is critical for productive septum formation in mycobacteria.	37525.400	0.007632928	103	0.048433885
Rv3793	Involved in the biosynthesis of the mycobacterial cell wall arabinan and resistance to ethambutol (EMB; dextro-2,2'-(ethylenediimino)-DI-1-butanol). Polymerizes arabinose into the arabinan of arabiogalactan [catalytic activity: UDP-L-arabinose + indol-3-ylacetyl-myo-inositol = UDP + indol-3-ylacetyl-myo-inositol L-arabinoside].	36733.140	0.06872068	136	0.048790414
Rv2111c	Involved in proteasomal degradation. Covalently binds to protein substrates.	35411.240	0.004444169	37	0.047942717
Rv1587c	unknown	35068.133	0.01124408	54	0.048124634
Rv1476	unknown	34379.210	0.09248264	148	0.048805457
Rv0002	DNA polymerase III is a complex, multichain enzyme responsible for most of the replicative synthesis in bacteria. This DNA polymerase also exhibits 3' to 5' exonuclease activity. The beta chain is required for initiation of replication once it is clamped onto DNA, it slides freely (bidirectional and ATP-independent) along duplex DNA [catalytic activity: N deoxynucleoside triphosphate = N diphosphate + {DNA}N].	33375.504	0.003620375	90	0.048279647
Rv3794	Involved in the biosynthesis of the mycobacterial cell wall arabinan and resistance to ethambutol (EMB; dextro-2,2'-(ethylenediimino)-DI-1-butanol). Polymerizes arabinose into the arabinan of arabiogalactan [catalytic activity: UDP-L-arabinose + indol-3-ylacetyl-myo-inositol = UDP + indol-3-ylacetyl-myo-inositol L-arabinoside].	31967.246	0.057621565	115	0.04876095

Rv Number	Function	Betweenness	Eigenvector	Degree	Closeness
Rv0883c	unknown	31779.750	0.05078291	112	0.04873993
Rv0430	unknown	31365.375	0.022703279	92	0.04852714
Rv3921c	unknown	29409.086	0.006799817	68	0.048375268
Rv0412c	unknown	29090.332	0.045916274	95	0.048480764
Rv3244c	unknown	28529.008	0.021864709	64	0.048388872
Rv3597c	Dominant T-cell antigen and possibly stimulates lymphoproliferation. Has DNA-bridging activity	27758.719	0.01894281	73	0.04840604
Rv1547	DNA polymerase III is a complex, multichain enzyme responsible for most of the replicative synthesis in bacteria. This DNA polymerase also exhibits 3' to 5' exonuclease activity. The alpha chain is the DNA polymerase [catalytic activity: N deoxynucleoside triphosphate = N diphosphate + {DNA}(N)].	26895.518	0.004110165	75	0.04831502
Rv0732	Essential for protein export. Interacts with SECA|Rv3240c and SECE|Rv0638 to allow the translocation of proteins across the plasma membrane, by forming part of a channel.	26715.15	0.011198077	111	0.04828495
Rv2751	unknown	26543.229	1.06E-04	12	0.046424046
Rv0088	unknown	26082.717	0.02309477	73	0.0482879
Rv2418c	unknown	24712.111	0.059896745	116	0.048605822
Rv1837c	Involved in glyoxylate bypass (second step), an alternative to the tricarboxylic acid cycle [catalytic activity: L-malate + CoA = acetyl-CoA + H(2)O + glyoxylate]	24302.383	0.002012468	47	0.04793749
Rv2186c	unknown	24297.744	0.08629693	148	0.04880907
Rv0467	Involved in glyoxylate bypass (at the first step), an alternative to the tricarboxylic acid cycle (in bacteria, plants, and fungi) [catalytic activity: isocitrate = succinate + glyoxylate]. Involved in the persistence in the host.	23928.854	0.002939687	53	0.048157435
Rv3490	Involved in osmoregulatory trehalose biosynthesis. Mycobacteria can produce trehalose from glucose 6-phosphate and UDP-glucose (the OtsA-OtsB pathway) from glycogen-like alpha(1-->4)-linked glucose polymers (the TreY-TreZ pathway) and from maltose (the TreS pathway) [catalytic activity: UDP-glucose + D-glucose 6-phosphate = UDP + alpha,alpha-trehalose 6-phosphate].	23777.691	7.30E-04	30	0.04769481
Rv1415	Involved in riboflavin biosynthesis [catalytic activity: GTP + 3 H(2)O = formate + 2,5-diamino-6-hydroxy-4-(5-phosphoribosylamino)pyrimidine + diphosphate].	23617.264	0.002809845	57	0.048159193
Rv0290	unknown	23576.32	0.0738134	128	0.048615973

Rv Number	Function	Betweenness	Eigenvector	Degree	Closeness
Rv2476c	Catabolic glutdh involved in the utilization of glutamate and other amino acids of the glutamate family. Generates 2-oxoglutarate from L-glutamate [catalytic activity: L-glutamate + H(2)O + NAD(+) = 2-oxoglutarate + NH(3) + NADH].	23372.959	0.001772863	60	0.048072018
Rv0236c	Involved in the biosynthesis of the mycobacterial cell wall arabinan	23372.916	0.03596256	77	0.04851048
Rv0260c	Could be involved in transcriptional mechanism.	23129.807	0.002362427	48	0.04810767
Rv2219	unknown	23090.479	0.031297263	101	0.04856049
Rv3240c	Involved in protein export. Interacts with the SECY/SECE subunits. SECA has a central role in coupling the hydrolysis of ATP to the transfer of PRE-secretory periplasmic and outer membrane proteins across the membrane.	22790.56	0.003690754	64	0.048158605
Rv0541c	unknown	22700.352	0.021116229	65	0.04832682
Rv2534c	Involved in peptide bond synthesis. Stimulate efficient translation and peptide-bond synthesis on native or reconstituted 70S ribosomes in vitro. Probably functions indirectly by altering the affinity of the ribosome for aminoacyl-tRNA, thus increasing their reactivity as acceptors for peptidyl transferase.	22585.900	0.010742613	123	0.04837822
Rv0227c	unknown	22294.797	0.07554747	111	0.048643466
Rv3780	unknown	22001.602	0.016156351	70	0.04835221
Rv3201c	Has both ATPase and helicase activities	21984.812	0.006370678	39	0.048035834
Rv1382	unknown	21967.104	0.03207123	87	0.048534878
Rv2868c	unknown	21919.65	0.00296851	61	0.048131075
Rv1738	unknown	21804.758	0.01357029	66	0.048354577
Rv0707	This protein is involved in the binding of initiator met-tRNA	21619.605	0.012702598	138	0.04832092
Rv2093c	Involved in proteins export: required for correct localization of precursor proteins bearing signal peptides with the twin arginine conserved motif S/T-R-R-X-F-L-K. This sec-independent pathway is termed tat for twin-arginine translocation system. This system mainly transports proteins with bound cofactors that require folding prior to export.	21413.893	0.00274251	52	0.048146885
Rv0585c	unknown	21225.855	0.020321168	66	0.048333902
Rv2553c	unknown	21154.973	0.004895513	53	0.04829615
Rv0666	unknown	21065.979	0.001632791	16	0.047455717
Rv2507	unknown	20701.37	0.0505186	93	0.048471857

Rv Number	Function	Betweenness	Eigenvector	Degree	Closeness
Rv3658c	unknown	20682.938	0.043272506	87	0.04855036
Rv0216	unknown	20616.273	0.019398635	57	0.048212588
Rv0431	unknown	20463.848	0.020625127	42	0.048121125
Rv1797	unknown	20461.07	0.0607339	102	0.048521783
Rv1602	Histidine biosynthesis pathway (fifth step). Catalyzes an amidotransferase reaction that generates imidazole-glycerol phosphate and 5-aminoimidazol-4-carboxamide ribonucleotide, which is used for purine synthesis.	20447.479	0.002363527	49	0.048105914
Rv0511	Possibly involved in the biosynthesis of siroheme and cobalamin [catalytic activity: 2 S-adenosyl-L-methionine + uroporphyrin III = 2 S-adenosyl-L-homocysteine + sirohydrochlorin].	20102.896	0.00593232	54	0.04820789
Rv1794	unknown	19814.705	0.05045415	88	0.048460577
Rv2700	unknown	19611.584	0.080525346	132	0.048723727
Rv2097c	Ligates deamidated pup to proteasomal substrate proteins. Requires hydrolysis of ATP to ADP.	19334.732	0.00727773	57	0.048266105
Rv2647	unknown	19166.783	0.001201815	12	0.047156546
Rv1044	unknown	19090.459	0.003662011	25	0.04765575
Rv0292	unknown	18595.605	0.05775365	93	0.04844871
Rv2813	unknown	18302.37	0.001109125	28	0.04772127
Rv0355c	unknown	18283.836	0.006414217	41	0.047984574
Rv1159	Polyprenol-phosphate-mannose dependent mannosyltransferase involved in phosphatidylinositol mannoside synthesis	17943.375	0.007109766	41	0.047953174
Rv2492	unknown	17856.270	0.002281833	27	0.047683317
Rv2552c	Possibly involved at the fourth step in the biosynthesis of chorismate within the biosynthesis of aromatic amino acids (the shikimate pathway) [catalytic activity: shikimate + NADP(+) = 5-dehydroshikimate + NADPH].	17841.855	0.002641129	61	0.04819556
Rv1611	Tryptophan biosynthesis pathway (fourth step) [catalytic activity: 1-(2-carboxyphenylamino)-1-deoxy-D-ribulose 5-phosphate = 1-(indol-3-YL)glycerol 3-phosphate + CO(2) + H(2)O.]	17798.758	0.002891334	57	0.048196144

Rv Number	Function	Betweenness	Eigenvector	Degree	Closeness
Rv2847c	Involved in the biosynthesis of siroheme and cobalamin [catalytic activity: 2 S-adenosyl-L-methionine + uroporphyrin III = 2 S-adenosyl-L-homocysteine + sirohydrochlorin]. SAM-dependent methyl transferase that methylates uroporphyrinogen III at position C-2 and C-7 to form precorrin-2 and then position C-12 or C-18 to form trimethylpyrrocorphin 2. It catalyzes also the conversion of precorrin-2 into siroheme (consisting of an oxidation and FE(2+) chelation).	17485.742	0.001709926	56	0.04795201
Rv2773c	Involved in biosynthesis of diaminopimelate and lysine from aspartate semialdehyde (at the second step) [catalytic activity: 2,3,4,5-tetrahydrodipicolinate + NAD(P)(+) = 2,3-dihydrodipicolinate + NAD(P)H].	17351.877	0.001899245	52	0.048025925
Rv0249c	Could be involved in interconversion of fumarate and succinate (aerobic respiration). This hydrophobic component may be required to anchor the catalytic components of the succinate dehydrogenase complex to the cytoplasmic membrane.	17255.219	0.020175789	61	0.04831443
Rv3792	Involved in the biosynthesis of the mycobacterial cell wall arabinan	17136.64	0.03662559	76	0.04834867
Rv2474c	unknown	17136.035	0.033315834	70	0.048409592
Rv1315	Involved in cell wall formation; peptidoglycan biosynthesis. Adds enolpyruvyl to UDP-N-acetylglucosamine [catalytic activity: phosphoenolpyruvate + UDP-N-acetyl-D- glucosamine = phosphate + UDP-N-acetyl-3-O-(1-carboxyvinyl)-D-glucosamine]	16958.688	0.003925464	66	0.04808779
Rv0955	unknown	16952.053	0.04612141	72	0.048436258
Rv0902c	Sensor part of the two component regulatory system PRRA/PRRB. Thought to be involved in the environmental adaptation, specifically in an early phase of the intracellular growth.	16640.244	0.021115992	42	0.04806676
Rv1402	Recognizes a specific hairpin sequence on PHIX SSDNA; this structure is then recognized and bound by proteins PRIB and PRIC. Formation of the primosome proceeds with the subsequent actions of DNAB, DNAC, DNAT and primase. PRIA then functions as a helicase within the primosome	16635.785	0.004491256	67	0.0481803
Rv1711	unknown	16528.084	0.001895638	59	0.048037585
Rv2455c	probably involved in cellular metabolism.	16318.104	0.00160155	54	0.04806034
Rv0658c	Supposedly involved in stationary-phase survival.	16228.710	6.80E-06	5	0.045444105
Rv1343c	unknown	16204.041	0.048564665	104	0.048619557
Rv0289	unknown	16198.578	0.075001664	108	0.048549764
Rv1854c	Transfer of electrons from NADH to the respiratory chain. The immediate electron acceptor for the enzyme is believed to be ubiquinone. Does not couple the redox reaction to proton translocation.	16198.030	0.001777024	37	0.04796422
Rv1024	unknown	16154.034	0.009733778	52	0.04823492

Rv Number	Function	Betweenness	Eigenvector	Degree	Closeness
Rv2172c	unknown	15838.416	0.070005916	106	0.048534878
Rv2980	unknown	15756.918	0.03168082	69	0.048333313
Rv2711	Transcriptional regulatory protein, iron-binding repressor of siderophore biosynthesis and iron uptake. Seems to regulate a variety of genes encoding a variety of proteins e.g. transporters, proteins involved in siderophore synthesis and iron storage, members of the PE/PPE family, enzymes involved in lipid metabolism, transcriptional regulatory proteins, etc. Also activator of BFRA|Rv1876 gene.	15661.199	0.005902891	41	0.04817209
Rv0651	Involved in translation mechanisms.	15638.877	0.012551642	118	0.04836226
Rv3843c	unknown	15604.246	0.049048502	76	0.048473045
Rv1128c	unknown	15471.695	0.029202195	73	0.048426773
Rv3805c	Involved in the biosynthesis of the mycobacterial cell wall arabinan	15423.152	0.03873028	70	0.048389465
Rv3601c	Involved in pantothenate biosynthesis [catalytic activity: L-aspartate = beta-alanine + CO(2)].	15341.172	0.00251573	49	0.048123464
Rv2112c	Deamidates the C-terminal glutamine of pup	15258.443	0.008601563	58	0.048300274
Rv2191	unknown	15224.334	0.008907985	45	0.048354577
Rv0383c	unknown	15151.17	0.026302978	65	0.048318557
Rv2518c	Involved in peptidoglycan synthesis. Catalyzes the formation of 3->3 crosslinks between peptidoglycan subunits.	14995.930	0.002818533	25	0.047732208
Rv3531c	unknown	14827.180	0.009694146	42	0.048041668
Rv2412	Involved in translation mechanisms. Binds directly to 16S ribosomal RNA.	14768.682	0.007476856	73	0.047918912
Rv2710	The sigma factor is an initiation factor that promotes attachment of the RNA polymerase to specific initiation sites and then is released. May control the regulons of stationary phase and general stress resistance. Seems to be regulated by sigh (Rv3223c product) and SIGE (Rv1221 product). Seems to regulate KATG|Rv1908c and the heat-shock response.	14727.146	0.004015876	53	0.04814454
Rv3660c	Possibly plays a regulatory role in celular differentiation.	14637.478	0.009687395	43	0.048086036
Rv3635	unknown	14474.549	0.049930908	77	0.048468295
Rv0736	Regulates negatively SIGL|Rv0735	14264.871	0.015362771	50	0.048300866
Rv2538c	Involved at the second step in the biosynthesis of chorismate within the biosynthesis of aromatic amino acids (the shikimate pathway) [catalytic activity: 7-phospho-3-deoxy-arabino-heptulosonate = 3-dehydroquinate + orthophosphate].	14206.770	0.002324578	65	0.0481065

Rv Number	Function	Betweenness	Eigenvector	Degree	Closeness
Rv1575	unknown	14179.283	0.002602459	17	0.04743353
Rv0164	unknown	14149.578	0.020210944	39	0.048143957
Rv1845c	Involved in transcriptional mechanism	14084.016	0.018668102	44	0.048171505
Rv3808c	Converts UDP-galactofuranose in cell wall galactan polymerization. Has UDP-Galf:beta-D-(1->5) and UDP-Galf:beta-D-(1->6) galactofuranosyltransferase activities.	13873.782	0.008149379	51	0.048013106
Rv2151c	This protein may be involved in septum formation.	13691.68	0.008896383	54	0.048147473
Rv0455c	unknown	13637.751	0.087514624	125	0.04874173
Rv2416c	Acetylation, substrate unknown. Involved in intracellular survival. Possibly associated with the cell surface and secreted. Modulates cytokine secretion by host immune cells.	13419.646	0.00771849	39	0.048289075
Rv0066c	Involved in the KREBS cycle [catalytic activity: isocitrate + NADP+ = 2-oxoglutarate + CO(2) + NADPH].	13193.767	8.49E-04	34	0.047693662
Rv2987c	Involved in leucine biosynthesis (at the second step) [catalytic activity: 3-isopropylmalate = 2-isopropylmaleate + H(2)O (also catalyses 2-isopropylmaleate + H(2)O = 3-hydroxy-4-methyl-3-carboxypentanone)].	13133.913	0.001755499	52	0.04798341
Rv2986c	This protein belongs to the histone like family of prokaryotic DNA-binding proteins which are capable of wrapping DNA to stabilize it, and prevent its denaturation under extreme environmental conditions.	13109.861	0.00315177	40	0.048122294
Rv3593	unknown	12906.409	0.04523719	57	0.04818793
Rv2444c	Thought to be involved in several cellular process.	12851.090	0.003267201	42	0.04797934
Rv1712	Catalyzes the transfer of a phosphate group from ATP to either CMP or dCMP to form CDP or dCDP and ADP [catalytic activity: ATP + CMP = ADP + CDP].	12815.694	0.002191681	62	0.048019513
Rv0286	unknown	12808.898	0.039533693	75	0.04826257
Rv2231c	Involved in cobalamin biosynthesis	12805.972	0.002218283	39	0.04781122
Rv2339	Thought to be involved in fatty acid transport.	12728.273	0.008924487	29	0.047922973
Rv2391	Catalyzes the reduction of sulfite to sulfide, in the biosynthesis of sulfur-containing amino acids and co-factors	12401.702	0.001206348	45	0.04787485
Rv1783	unknown	12333.329	0.010263415	46	0.048000872
Rv0051	Unknown	12332.349	0.026025003	65	0.04835044

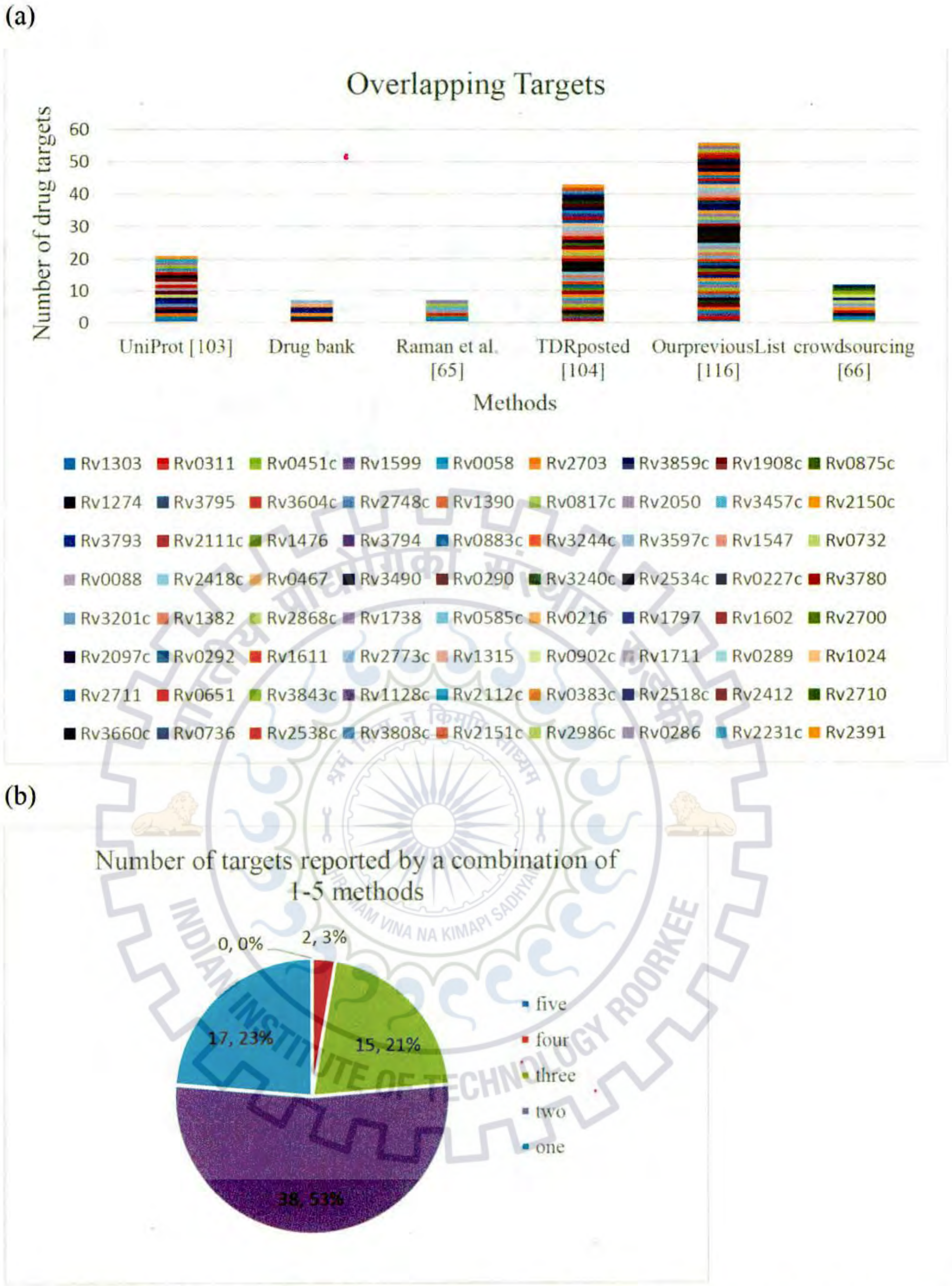


Figure 3.4 Details of proteins in the proposed list that have been reported by other methods. (a) The number of targets in the proposed list that have also been reported as potential drug targets by other methods. (b) Shows the number of targets from our proposed list that has been reported by all five, four, three, two and only one methods.

The resulted list proteins were assessed through a comparison with some of validated and potential targets. These targets were identified by using different computational and experimental methods. The dataset for this purpose was obtained by integrating manually curated targets from TDR, high confidence targets from UniProt, attractive targets obtained by Raman et al.[65] through a series of comprehensive filters, the potential drug target list identified in the previous chapter [116], non-redundant protein targets from DrugBank and targets identified through Crowd Sourcing [66]. The overlaps among these lists of drug targets and the proposed potential target list have been shown in Figure 3.4 (a). Based on this assessment, 43 proteins in the list were TDR validated targets, 6 of which were in the UniProt target list. An additional of 18 proteins in our list were overlapped with UniProt's list, 5 of which were also predicted by Raman et al. Raman et al.'s list contains 2 more proteins. From our previous report, 56 proteins were overlapped with the current candidates; some of them were already reported as potential targets by other methods. The comparison of list of targets proposed in this study with all of the approved, nutraceutical, illicit, investigational, withdrawn and experimental non-redundant protein targets of *Mycobacterium tuberculosis H37Rv* from DrugBank yields an overlap of 7 proteins. Potential drug targets from a comprehensive analysis of crowd sourcing have been also used for comparison [66]. Crowd sourcing is a new paradigm for interactome driven drug target identification in mycobacterium tuberculosis that has been carried out through extensive re-annotation and constructing a systems level protein interaction map of Mtb to identify the intended drug target candidates. Out of 137 proteins, 12 were reported as candidate targets of the pathogen by this method. Moreover, there are four known targets of existing anti-tubercular drugs within this set. They are Rv1908c (KatG) (ranked 12thin the proposed list), Rv3795 (EmbB) (ranked 15th), Rv3793 (ranked 25th) and Rv3794 (ranked 30th). Rv1908c (KatG) is a validated drug target of Isoniazid. Rv3793 and Rv3794 are target proteins of Ethambutol [68]. Rv3795 (EmbB) is a drug target for Rifampin, Isoniazid and Ethambutol. Hence, a total of 72 (52.6%) proteins from our proposed list have been previously predicted or reported to be drug targets by the stated methods. As it has been shown in Figure 3.4 (b), there is no target reported by all of the five methods. There are two proteins (Rv0058, Rv2150c) reported by four methods. 15 proteins were reported by three methods. There are 38 proteins reported by two methods and 17 proteins reported only by one method each.

The lists of top 20 proteins according to each of the four centrality measures have been obtained. From these lists, 10 of the proteins are found to be common and they are listed in Table 3.2. It is hypothesised that these proteins are better targets since they have been identified in higher ranks of the four centrality measures of the interactome network.

Table 3.2 Proteins in the top 20 of all of the four centrality measures

protein	Functional Class	Network Centrality Scores				PDB
		Betweenness	Eigenvector	Degree	Closeness	
Rv1303	cell wall and cell processes	115190.99	0.1066408	207	0.0490584	
Rv3019c	cell wall and cell processes	71224.33	0.08889492	167	0.0488181	3H6P
Rv0311	conserved hypotheticals	67942.516	0.07391167	156	0.048861504	
Rv0556	cell wall and cell processes	66400.05	0.103606604	193	0.049006738	
Rv0451c	cell wall and cell processes	66230.336	0.09615834	171	0.04883437	2LW3
Rv0288	cell wall and cell processes	63475.453	0.09102694	170	0.048855472	2KG7
Rv0875c	cell wall and cell processes	49197.566	0.11939961	197	0.04895823	
Rv1274	cell wall and cell processes	46747.54	0.11324064	197	0.04900249	
Rv0817c	cell wall and cell processes	40821.99	0.11914455	196	0.048973985	
Rv0358	conserved hypotheticals	40230.57	0.07967692	154	0.048827738	

Additionally, potential drug targets of the pathogen that interact with the host have been identified to understand the infection mechanism using a dataset obtained from a computational prediction of *H. sapiens-Mycobacterium tuberculosis H37Rv* protein-protein interactions [101]. This dataset is thought as a golden dataset for host-pathogen interaction. As it has been shown in Table 3.3, 15 proteins from the proposed target lists interact with human. The reason for the presence of only few overlaps could be due to the fact that the host–pathogen interaction dataset is not comprehensive or the host interacting proteins do not necessarily have to be essential to the pathogen and non-homologous with human. Identifying proteins of the pathogen participated in the complex interplay with the host could significantly increase the reliability of the targets since these interactions are key factors in determining the outcome of the infection [117].

Table 3.3 Proposed targets which interact with the host

Protein	Functional Class	Betweenness	Eigenvector	Degree	Closeness
Rv1599	intermediary metabolism and respiration	58334.63	0.002786182	82	0.048330363
Rv1908c	virulence, detoxification, adaptation	50548.035	0.005637622	76	0.048456423
Rv2150c	cell wall and cell processes	37525.4	0.007632928	103	0.048433885
Rv3921c	cell wall and cell processes	29409.086	0.006799817	68	0.048375268
Rv0732	cell wall and cell processes	26715.15	0.011198077	111	0.04828495
Rv1415	intermediary metabolism and respiration	23617.264	0.002809845	57	0.048159193

Protein	Functional Class	Betweenness	Eigenvector	Degree	Closeness
Rv2534c	information pathways	22585.9	0.010742613	123	0.04837822
Rv2553c	cell wall and cell processes	21154.973	0.004895513	53	0.04829615
Rv1602	intermediary metabolism and respiration	20447.479	0.002363527	49	0.048105914
Rv1611	intermediary metabolism and respiration	17798.758	0.002891334	57	0.048196144
Rv2455c	intermediary metabolism and respiration	16318.1045	0.00160155	54	0.04806034
Rv3601c	intermediary metabolism and respiration	15341.172	0.00251573	49	0.048123464
Rv2538c	intermediary metabolism and respiration	14206.7705	0.002324578	65	0.0481065
Rv2987c	intermediary metabolism and respiration	13133.913	0.001755499	52	0.04798341
Rv1712	intermediary metabolism and respiration	12815.694	0.002191681	62	0.048019513

Further, the study of *Mycobacterium tuberculosis* virulence is another path which has got a lot of attention in the design of drugs with a new mechanism of action, the production of modern concepts and tuberculosis treatment schemes [118]. Virulence factors have evolved as a response to the host immune reaction. In recent times, many *mycobacterial* virulence genes that are essential for the virulence of *Mycobacterium tuberculosis complex* (MTBC) species have been reported by a number of studies. Most of these genes either encode enzymes of several lipid pathways, cell surface proteins, regulators and proteins of signal transduction systems or involved in *mycobacterial* survival inside the aggressive microenvironment of the host macrophages. We took a compiled list of virulence genes from Forrellad et al. (2013) and tried to observe the overlap with our proposed potential targets. It has been found out that five genes from the proposed potential target list are also reported as virulence genes [118]. These genes have been shown in Table 3.4.

Table 3.4 Genes reported as virulence factors

Gene name	Rv Number	Functional Class	Betweenness	Eigenvector	Degree	Closeness
sigA	Rv2703	information pathways	56665.156	0.004549677	80	0.04838473
katG	Rv1908c	virulence, detoxification, adaptation	50548.035	0.005637622	76	0.04845642
icl1	Rv0467	intermediary metabolism and respiration	23928.854	0.002939687	53	0.04815743
pafA	Rv2097c	intermediary metabolism and respiration	19334.732	0.00727773	57	0.04826610
ideR	Rv2711	regulatory proteins	15661.199	0.005902891	41	0.04817209

3.3.4 Structural Assessment

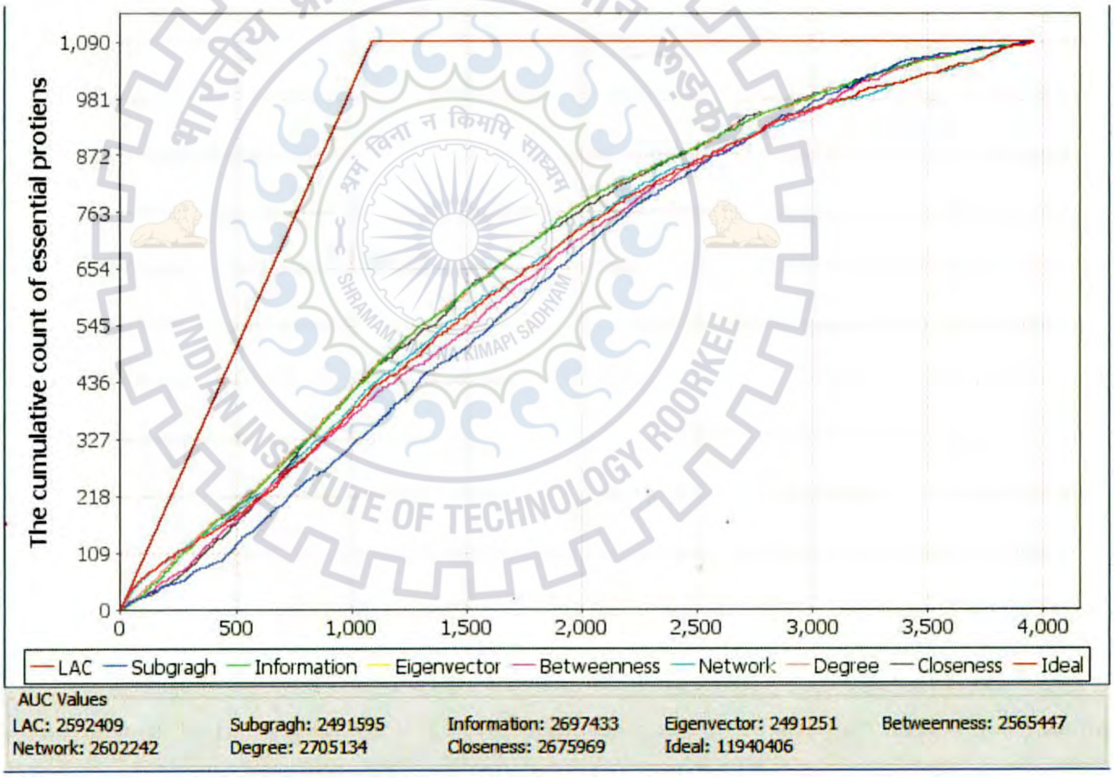
One of the main criteria which increase the targetablity of the prioritised lists of proteins is the availability of three-dimensional structures. The Protein Data Bank (PDB) is freely accessible and the main worldwide repository for the three-dimensional structural data of biological macromolecules such as proteins and nucleic acids which is typically obtained by X-ray crystallography or NMR spectroscopy and submitted by biologists and biochemists from around the world [119]. We checked the availability of solved structures of the identified potential lists of targets and out of 137 proteins from our proposed target list, 28 were successfully mapped to 82 structures from PDB which is approximated to 20.44%. This list has also been shown in Table 3.5 including the corresponding centrality measure values and PDB IDs of structures. Even though, experimentally solved structures are the primary preference in terms of reliability, identification of three-dimensional structures of protein through theoretically calculated homology models is a common practice which minimizes the problem of unavailability of structures.

Table 3.5 Sorted lists of proteins proposed as potential drug targets which have solved three-dimensional structure

Protein	Network Centrality Scores				PDB
	Betweenness	Eigenvector	Degree	Closeness	
Rv3019c	71224.33	0.08889492	167	0.0488181	3H6P;
Rv0451c	66230.336	0.09615834	171	0.0488344	2LW3;
Rv0288	63475.453	0.09102694	170	0.0488555	2KG7;
Rv0058	57680.18	0.00275039	62	0.0481486	2R5U;
Rv1908c	50548.035	0.00563762	76	0.0484564	1SFZ;1SJ2;2CCA;2CCD;4C50;4C51;
Rv2050	40088.305	0.01941196	91	0.0485105	2M4V;2M6P;
Rv2150c	37525.4	0.00763293	103	0.0484339	1RLU;1RQ2;1RQ7;2Q1X;2Q1Y;4KWE;
Rv3793	36733.14	0.06872068	136	0.0487904	3PTY;
Rv2111c	35411.24	0.00444417	37	0.0479427	3M91;3M9D;
Rv0002	33375.504	0.00362037	90	0.0482796	3P16;3RB9;
Rv3597c	27758.719	0.01894281	73	0.048406	2KNG;4E1P;4E1R;
Rv1837c	24302.383	0.00201247	47	0.0479375	2GQ3;3S9I;3S9Z;3SAD;3SAZ;3SB0;
Rv0467	23928.854	0.00293969	53	0.0481574	1F61;1F8I;1F8M;
Rv3240c	22790.56	0.00369075	64	0.0481586	1NKT;1NL3;
Rv0216	20616.273	0.01939864	57	0.0482126	2BI0;
Rv1611	17798.758	0.00289133	57	0.0481961	3QJA;3T40;3T44;3T55;3T78;4FB7;
Rv2773c	17351.877	0.00189925	52	0.0480259	1C3V;1P9L;1YL5;1YL6;1YL7;
Rv0902c	16640.244	0.02111599	42	0.0480668	1YS3;1YSR;
Rv2711	15661.199	0.00590289	41	0.0481721	1B1B;1FX7;1U8R;2ISY;2ISZ;2IT0;
Rv3601c	15341.172	0.00251573	49	0.0481235	2C45;
Rv2518c	14995.93	0.00281853	25	0.0477322	3VYN;3VYO;3VYP;4GSQ;4GSR;4GSU;4HU2;4HUC;
Rv0736	14264.871	0.01536277	50	0.0483009	3HUG;
Rv2538c	14206.7705	0.00232458	65	0.0481065	3QBD;3QBE;
Rv3808c	13873.782	0.00814938	51	0.0480131	4FIX;4FIY;
Rv2416c	13419.646	0.00771849	39	0.0482891	3R1K;3RYO;3SXO;3UY5;
Rv2987c	13133.913	0.0017555	52	0.0479834	3H5E;3H5H;3H5J;
Rv2986c	13109.861	0.00315177	40	0.0481223	4DKY;4PT4;
Rv2391	12401.702	0.00120635	45	0.0478749	1ZJ8;1ZJ9;

3.4 Assessment of the Method

It would be ideal to have standard validation data in order to assess the performance of the four centrality measures used in this analysis but it is not readily available. The list of essential proteins obtained through a comparative analysis has been used as a test data. Since the main objective of centrality measures in a network is to identify the proteins which are important in the interaction network, taking this data for evaluation seems reasonable. The four centrality measures used in this analysis were compared with other typical centrality measures: Local Average Connectivity-based method (LAC), Network Centrality (NC), Sub-graph Centrality (SC) and Information Centrality (IC). As it can be seen in the jack-knife line chart (Figure 3.5), there are no huge differences among the eight centrality measures with the AUC value of degree centrality the highest of all. Information and closeness centralities ranked second and third, respectively.



AUC Values
LAC: 2592409 Subgragh: 2491595 Information: 2697433 Eigenvector: 2491251 Betweenness: 2565447
Network: 2602242 Degree: 2705134 Closeness: 2675969 Ideal: 11940406

Figure 3.5 Jack-knife line chart of eight centrality measures. The cumulative count of essential proteins of eight different centrality measures have been shown to assess the performance of the four centrality measures used for this analysis

Another testing data, including validated drug targets, intersection of high confidence targets from UniProt and attractive targets from Raman et al. [65] were identified. This list contains 47 proteins. Then, the eight centrality measures were compared in terms of the average rank of the drug targets in which lower average rank indicates better performance. The absolute count of drug targets in 1% of all candidate proteins (practically in the top 40 proteins), in the top 5% (practically in the top 198 proteins), in the top 10% (practically in the top 396 proteins), in the top 15% (practically in the top 594 proteins) and in the top 20% (practically in the top 792 proteins) among all candidate were reported (Table 3.6). For instance, in the top 1%, betweenness and closeness centrality identified 2 drug targets each while the others found 1. Eigenvector identified joint maximum potential targets in all of top 5%, 10%, 15% and 20%. We took up to 20% for comparison because these proteins found near the center of gravity values.

Table 3.6 Number of drug targets and its average position among different methods in top 1%, 5%, 10%, 15% and 20% of the candidate proteins list

Method	**1%(40)**	**5 %(198)**	**10%(396)**	**15 %(594)**	**20 %(792)**
LAC	0	6	9	12	15
Subgraph	1	8	13	19	23
Information	1	6	9	13	17
Network	1	6	9	12	15
Eigenvector	1	8	13	19	23
Betweenness	2	7	10	14	18
Degree	1	6	9	13	17
Closeness	2	5	10	15	22

3.5 Conclusion

In this study, we have identified a list of proteins which could be an attractive and reliable target for *Mycobacterium tuberculosis H37Rv* through a comprehensive analysis of comparative genome and network centrality measures of protein-protein interaction network. The comparative genome analysis has helped in identifying those lists of proteins which are essential for the survival and growth of the pathogen to increase success rate of drugs to be designed. It was also useful in filtering out those proteins which are absent in human to eliminate all those with a risk of causing host toxicity. In traditional drug discovery the side effect or drug safety has normally been addressed by making modification on the drug molecule but systematic way of dealing with this problem at the drug target identification phase in the modern rational drug discovery process seems to be more effective [44]. The refined lists of proteins were then prioritized by using network centrality measures where proteins that found at the centre of gravity of interactome network were proposed as a final list of potential targets of the pathogen. Proteins found at the centre of gravity of the disease specific protein-protein interaction network are hypothesised to be more important proteins in the pathogen and hence more likely to be attractive targets. The comparison of these lists of targets with some of known drug targets as well as potential targets predicted by using different computational and experimental methods revealed that about half of them have been previously predicted or reported to be potential drug targets for *Mycobacterium tuberculosis H37Rv*. The structural assessment of these proteins has also showed that some proteins found to have experimentally solved three-dimensional structure. In general, we believe that this comprehensive analysis will have significant contribution in providing an important input for the experimental study of developing new antibiotics for infamous *Mycobacterium tuberculosis H37Rv*.

CHAPTER FOUR

Maximum flow approach to prioritize potential drug targets of *Mycobacterium tuberculosis H37Rv* from protein-protein interaction network

4.1 Introduction

Lists of potential drug targets of human diseases have been identified using computational methods such as structure-based, network-based and integrated approaches [46,65, 104]. In spite of their contributions, each computational drug target identification method has well documented limitations. In approaches that employ global network centrality measures such as closeness/betweenness to discover new drug targets from molecular interaction networks, a node with higher global centrality value would be considered as initial candidates for drug targets [120]. These methods are based on shortest paths. The non-shortest paths are not considered even though they may be important in spreading information among interacting biological entities in the cellular network. However, the shortest path length in a biological network is typically very small and most of the time there will be additional "relatedness" between two interacting nodes due to the small world property of biological networks [121, 122]. This is the common critic with respect to the global network centrality measures but there are some counter arguments. One of which is that shortest paths yield a higher coverage than observed directly neighbours locally from protein-protein interaction network. Moreover, it has been hypothesised that shortest paths are the most feasible paths that can be taken by proteins to communicate with each other [44]. Measure of betweenness centrality based on random walks has been tried and suggested as improved versions of these measures [123]. Even though this approach tries to incorporate all paths among nodes, it still gives more weight to the short paths.

Computational methods that have been used in the identification of drug targets for drug-resistance TB did not consider the influence of newly identified targets to drug resistance genes of First Line Drugs (FLDs) and Second Line Drugs (SLDs). There are some efforts to understand emergence of resistance mechanisms from wholistic system perspective in which co-targets have been suggested to prevent the emergence of drug-resistance [44, 68, 69]. Co-targets are proposed to be used in a systematically designed combination with primary targets. The main objective is to inhibit the primary target and co-target simultaneously so as cellular interaction between the primary target and resistance mechanisms can be disrupted. This could be an effective method to prevent the emergence of drug resistance. In spite of the importance of co-targets to tackle the emergence of drug resistance, it is not helpful in treating already developed drug-resistance TB.

That is one of the main reasons why we need new primary targets than co-targets for the development of efficient therapeutics for the treatment of already developed drug-resistance TB.

In this analysis, maximum flow approach has been used as a main method to further prioritize drug targets of *Mycobacterium tuberculosis H37Rv* from weighted protein-protein interaction network .The weighted protein-protein interaction network of the pathogen was constructed by using a dataset from STRING database where only interactions with combined score value greater than or equal to 770 were considered. From the generated protein-protein interaction network of the pathogen, maximum flows of the drug target proteins to resistance genes were computed. Then, drug targets were prioritized based on their maximum flow values. Since this approach is based on the flow, it is not expected to be affected by biasness towards shortest paths like the common global centrality measures. More importantly, the inhibition of a protein which has a maximum flow to the resistance genes of the existing drugs is expected to disrupt the communication to these genes. So, this further prioritization is believed to be a rational approach to deal with the problem of resistance at the initial stage of drug discovery process.

4.2 Materials and Method

A set of highly reliable potential drug target set P and disease resistance genes set G of *Mycobacterium tuberculosis H37Rv* have been taken as input. The potential drug target set P consists of a prioritized list of 137 proteins that has been identified from the comparative genome and network centrality analysis in chapter 2. The drug resistance genes list G is a curated list of 82 genes involved in both intrinsic and extrinsic drug resistance mechanisms of *Mycobacterium tuberculosis H37Rv* identified from two published papers [65, 68]. They have scanned the available biological literature to obtain information about associations of individual proteins with drug-resistance and verified manually to include in the drug-resistance list.

Subsequently, weighted protein-protein interaction network of the pathogen was constructed by using a dataset retrieved from Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) database [96]. A combined score is assigned for every pair of protein-protein association in the database [109]. This score is computed by combining the probabilities from several pieces of evidence and correcting for the probability of randomly observing an interaction. The higher combined score value for interacting pair of proteins means the interaction is being supported by several pieces of evidence. It has been shown in a recent comprehensive study that the protein-protein interactions of *Mycobacterium tuberculosis H37Rv* generated from STRING are of low quality consisting of a significant number of false positives and false negatives [110]. However, the study also indicates that a subset of this dataset (interactions with combined score $\geq$ 770) is more reliable with greater portion of it having correlated gene expression profiles and coherent informative Gene Ontology (GO) term annotations in both interaction partners. This subset of interaction dataset has been used in this analysis. The combined score of the pair of interacting proteins has been assigned as a weight of interaction [124]. It is hypothesized that a higher weight for a pair of interacting proteins implies more flow. The weight zero is assumed for non-interacting pair of proteins as no flow can pass through.

The main focus of this analysis is to further prioritise potential drug targets of *Mycobacterium tuberculosis H37Rv* based on their maximum flow to resistance genes. This has been done through identifying the maximum flow between the candidate proteins P and disease genes G. Maximizing the flow from the candidate proteins to the resistance genes seems to be reasonable since understanding the efficient communications in the biological networks can be helpful to design treatment mechanisms for the problem of resistance in a systematic and rational way. The following are the constraints applied to specify the problem into the classical maximum flow problem [46]:

Constraints 1: The protein-protein interaction is a bi-directional edge.
In the protein-protein interaction network, flow direction is one of the most important features. However, almost all the outcomes of current high-throughput techniques for protein-protein interactions mapping are usually supposed to be non-directional. The protein-protein interaction in this analysis has been considered as bi-directional edge.
Constraints 2: A dummy sink node is created to connect all resistance genes with the capacity of these edges set to infinity.
Constraint 3: All candidate proteins are connected to a dummy source node and the capacity of these edges are set to infinity.

Since infinite value is not practical, the value greater than the maximum possible flow has been assigned.

Constraints 4: If a resistance gene is also a candidate protein, to avoid unfair advantage, it has only been connected to dummy source node but not to the dummy sink node.

Let V be a set of nodes representing proteins, E be a set of edges representing interactions between proteins, and the nodes *s* and *t* be the source and sink nodes respectively. Further, let *c* be a nonnegative capacity of an edge. Then, for a weighted interaction network G = (V, E, *s*, *t*, *c*), the size is denoted as n=|V| and m=|E|. The *flow f* is a real valued function on the edges and the excess value $e_f(v)$ is the difference between the incoming and outgoing flows to a node under consideration. A node v is said to be active if $v \in V - \{s,t\}, d(v) < n$ and $e_f(v) > 0$. The edge (v,w) is admissible if $d(v) = d(w) + 1$. The *flow f* satisfies the following three constraints [81, 87]:

Capacity constraint: $f(v,u) \leq c(v,u) \quad for\ all\ (v,u) \in V \times V$ (1)

Skew symmetry constraint: $f(v,u) = -f(u,v) \quad for\ all\ (v,u) \in V \times V$ (2)

Flow conservation constraint: $\sum f(u,v) = 0 \quad for\ all\ (u,v) \in V - \{s,t\}$ (3)

The problem here is to find the maximum flow from the dummy source node to dummy sink node. It can be computed by manipulating the *preflow f* on the network satisfying constraints (1) - (3).

The following initializations have been made [87]:

The initial value of the *flow* $f(v,u)$ is set to zero for all $(v,u) \in (V - \{s\}) \times (V - \{s\})$ and $f(s,v)\ set\ to\ c(s,v)$ for all $v \in V$.

The distance of a node $d(v)$ to the sink node t is n for $v = s$ and zero for all $v \in (V - \{s\})$.

The excess flow e_f for the source node is the sum of capacities of all of its edges and zero for all vertices $v \in (V - \{s,t\})$.

Once the initialization is complete, a repeated *push* and *relabel* operations have been performed on active nodes starting from the provided source node until there are no more active nodes or the edges are *saturated.*

The *Push* (v,w) and *Relabel* (v) operations:

Push (v,w).
Applicability: v is active and (v,w) is admissible
Action: send $\delta = min\ (e_f(v), u_f)$ units of flow from v to w

Relabel (v)
Applicability: v is active and *push* (v,w) can't be applied

Action: replace $d(v)$ by $min\ \{d(w) + 1 | (v, w) \in E_f\ or\ by\ n\ otherwise\}$

An edge is said to be *saturated* if the flow on it can't be increased without violating the capacity constraint and *residual* otherwise [81]. The *residual capacity* u_f of an edge (v, w) is the amount by which the flow can be increased. The *push* and *relabel* operations modify the *preflow f* and labelling *d*. A push from *v* to *w* increases $f\ (v, w)$ and $e_f(w)$ by $\delta = min\ (e_f(v), u_f)$, and decreases *f (w, v)* and $e_f(v)$ by the same amount. A relabeling of *v* sets its label to the largest value allowed by the valid labeling constraints. FIFO algorithm has been used to maintain the set of active nodes in which the front node is always selected for discharging and the newly active node is added to the rear of the queue [81]. *Gap relabeling* has been used as a distance relabeling heuristic which is based on the following observation [70]. If we have an integer *g* and $0 < g < n$. Suppose at a certain stage of the algorithm there are no nodes with distance label *g* but there are nodes *v* with $g < d\ (v) < n$. This implies that the sink is not reachable from any of these nodes. Therefore, the labels of such nodes may be increased to *n* (note that these nodes will never be active).

The implementation contains a comprehensive iterative operation called discharge operation. It consists of a repeated *push*, *relabel* and *gap relabel* operations.

```
Discharge $(v)$
Applicability: $v$ is active.
Action: let $(v, w)$ be the current edge of $v$;
    $end - of - list \leftarrow \boldsymbol{false}$;
    repeat
        if $(v, w)$ is admissible then push$(v, w)$
        else
        if $(v, w)$ is not the last edge on the edge list of $v$ then
            replace $(v, w)$ as the current edge of $v$ by the next edge on the list
        else begin
            make the first edge on the edge list of $v$ the current edge;
            $end - of - list \leftarrow \boldsymbol{true}$;
            end;
    until $e_f(v) = 0$ or end-of-list;
    if end-of-list then gap relabel $(v)$ / relabel $(v)$;
```

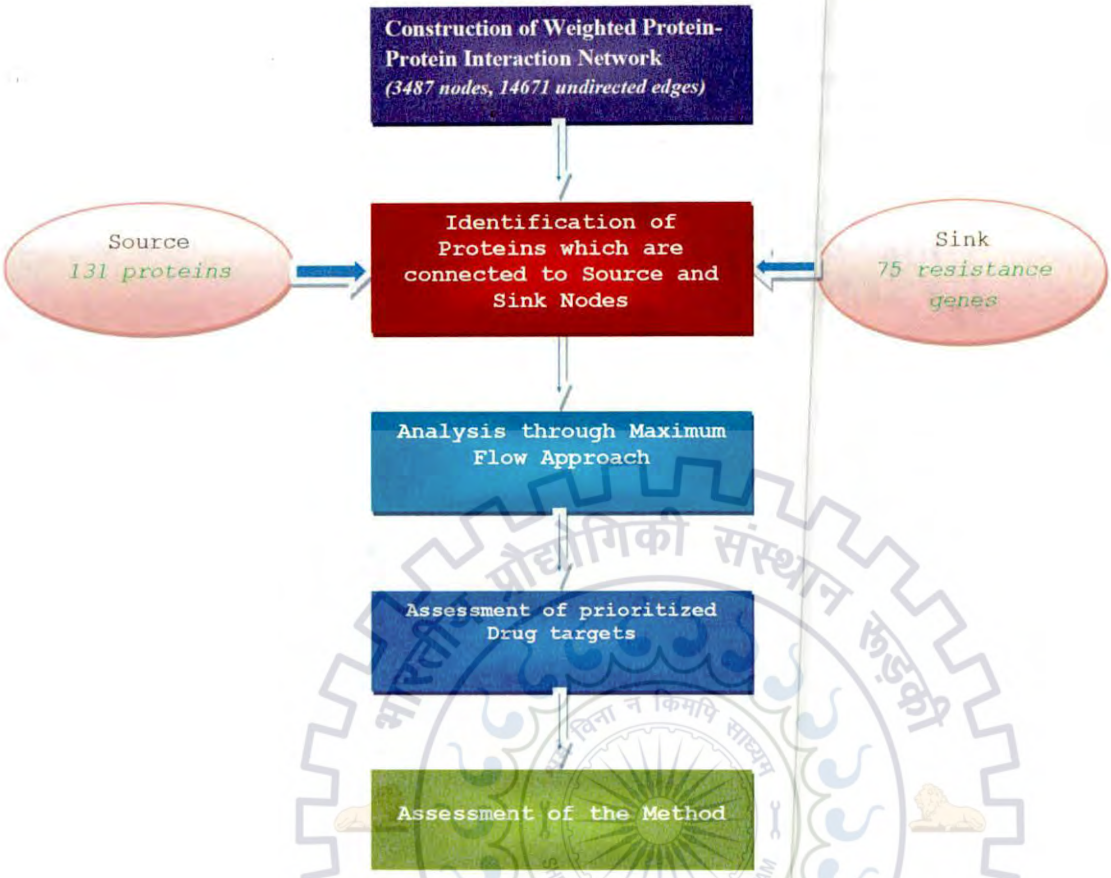

Construction of Weighted Protein-Protein Interaction Network
(3487 nodes, 14671 undirected edges)
Source
131 proteins
Identification of Proteins which are connected to Source and Sink Nodes
Sink
75 resistance genes
Analysis through Maximum Flow Approach
Assessment of prioritized Drug targets
Assessment of the Method

Figure 4.1 Progression of the experiments

Cytoscape 3.0.2 was used for the generation of statistical properties of the network [89]. The network centrality measures were computed using CytoNCA, a plug-in of Cytoscape [90]. The maximum flow approach was implemented using adjacency list of First In First Out (FIFO) *push-relabel* maximum flow with *gap relabeling* heuristic. For the purpose of proof of correctness, a graph obtained from an analysis by Schroeder et al. (2004) has been used [125]. The progression of experiments of this analysis has been shown in Figure 4.1.

4.3 Results and Discussion

Discovery of new drugs of TB is as one approach to deal with the problem of drug-resistance more systematically [29]. In this study, potential drug targets of *Mycobacterium tuberculosis H37Rv* have been further prioritized based on their maximum flow to the resistance genes to identify targets in which their inhibition would disrupt communications to the resistance genes in the protein-protein interaction network. The initial input set of potential drug targets of the pathogen, consisting of a list of 137 proteins, is obtained from systematically integrated comparative genome and network centrality analysis. These proteins are believed to be more reliable targets since they are essential for the survival and growth of the pathogen, non-homologous to human and found within the close neighbourhood of the centre of gravity of the protein-protein interaction network of the pathogen.

There are four mechanisms of resistance for *Mycobacterium tuberculosis H37Rv*: efflux pumps, target-modification, DNA replication and Horizontal Gene Transfer (HGT) [44]. A curated list of 82 proteins involved in these drug resistance mechanisms has also been taken as an input for the study. Then, protein-protein interaction network was constructed based on the retrieved associations from STRING [96]. Only interactions with a combined score value $\geq$ 770 were considered since they are verified as more reliable dataset containing less false negatives and false positives [110]. The resulting refined network contains 14,671 interactions among 3,487 proteins.

General statistical properties of the generated network have been shown in Table 4.1. Characteristic path length is an important property of a network indicating one protein's influence on another with number of intermediate reactions which would be useful to understand the efficiency of communication of biological information. The corresponding shortest path length distribution has been shown in Figure 4.2. Another important property of a network is *clustering coefficient*. The *clustering coefficient* of the resulted network is significantly higher than the *clustering coefficient* of a random graph with the same number of vertices (0.002). The degree distribution of the resulted network has also been shown in Figure 4.3. The network exhibits scale-free property since the degree distribution of proteins approximates a power law $p(k) = k^{-\gamma}$, with the degree exponent $\gamma \sim 1.753$. This means there are very small number of highly connected nodes called hubs and a vast majority of nodes with few connections.

Table 4.1 Statistical properties of the generated network

Parameter	Value
Number of nodes (n)	3485
Connected components	106
Network diameter	15
Average number of neighbours	8.420
Network density	0.002
Characteristic path length	5.655
Clustering coefficient	0.379

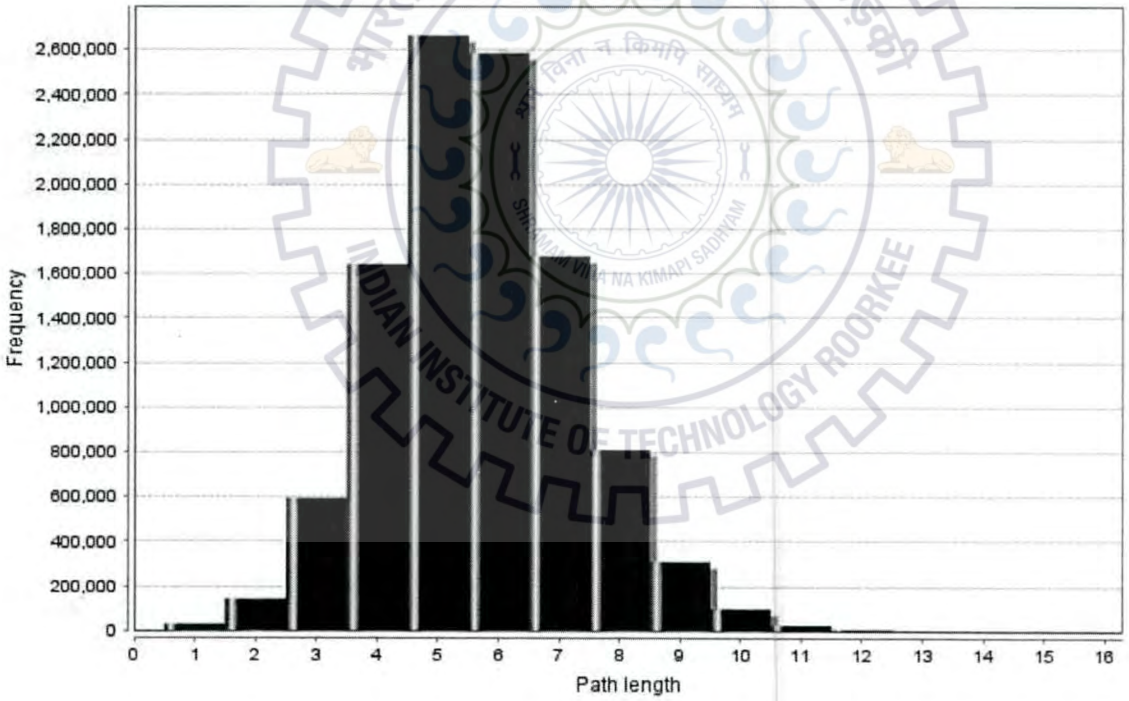


Figure 4.2 Shortest path length distributions

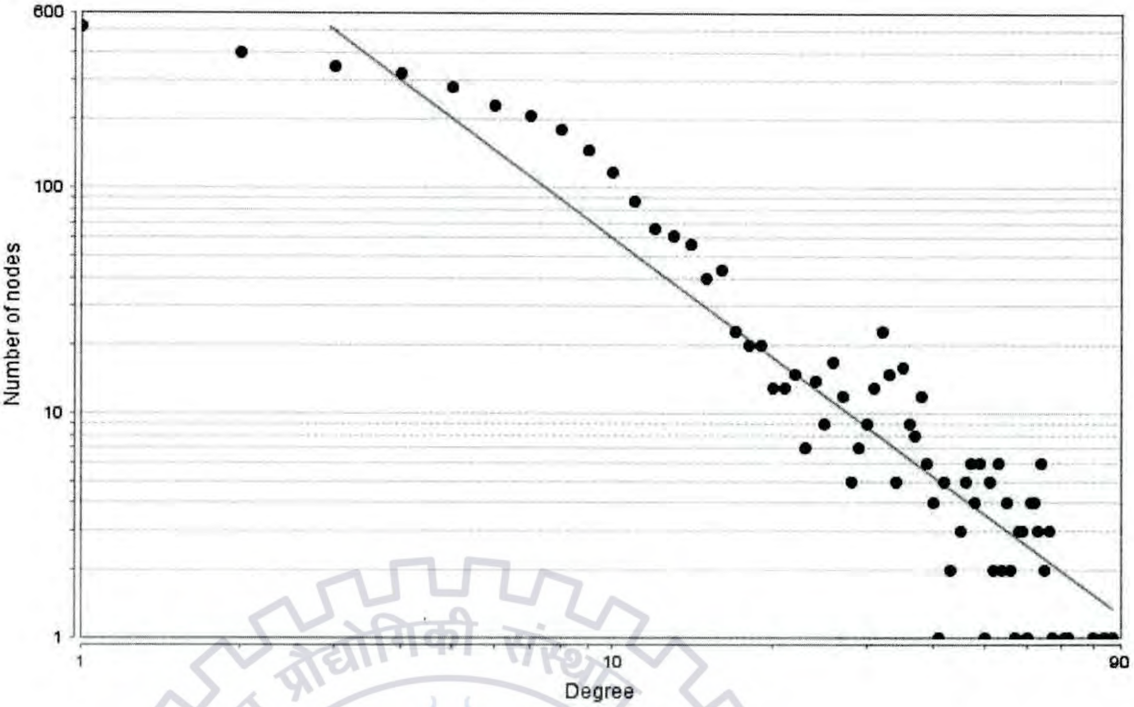


Figure 4.3 Node degree distributions

The source and sink nodes have been defined to find the maximum flow from the potential drug targets to the resistance genes and prioritize them accordingly. Out of the 137 proteins taken as potential drug target set, 131 proteins were found in the resulted protein-protein interaction network and these proteins were connected to the dummy source node based on constraint 3. From the curated list of 82 resistance genes, 78 were found in the constructed protein-protein interaction network. Three of them were also candidate drug targets and they have only been connected to dummy source node based on constraint 4. Then the remaining 75 proteins (Table 4.2) were connected to a dummy sink node based on constraint 2.

The maximum flows of each of 131 potential candidate proteins to the dummy sink node have been identified and they were sorted accordingly. It is hypothesize that flows are used to quantify structural and biochemical signal flows from the candidate proteins to the molecular components of the resistance machinery by which inhibition of the newly proposed targets are expected to have a better success in dealing with MDR and XDR TB. The list of potential drug targets has been shown in Table 4.3 which incorporates maximum flow value, functional category and pathway name of each candidate protein.

Table 4.2 Resistance Genes. A curated list of resistance genes involved in both intrinsic and extrinsic drug-resistance mechanisms of *Mycobacterium tuberculosis H37Rv* retrieved from literatures. The list contains protein coding genes found in generated proteome network but not on potential drug target set.

Resistance Mechanism	**Proteins in the Resistome**
Antibiotic efflux pumps	PstB (Rv0933), Rv2686c, Rv2687c, IniA (Rv0342), Rv3728, Rv2846c, Rv1877, Rv2459, Rv1410c, Rv1258c, Rv0783c, Rv1634, Rv0849
Hypothetical efflux pumps	Rv0191, Rv0037c, Rv2456c, Rv2994, Rv0262c, Rv1819c, Rv2209, Rv2477c, Rv2688c, Rv2938, Rv3361c
Antibiotic degrading enzymes	BlaC (Rv2068c)
Target-modifying enzymes	Erm37 (Rv1988), WhiB7 (Rv3197A)
SOS and related genes	DnaE2 (Rv3370c), RuvA (Rv2593c), RecA (Rv2737c), RecB (Rv0630c), RecC (Rv0631c), RecD (Rv0629c), PolA (Rv1629), LexA (Rv2720), Rv0818
Genes implicated in HGT	SecA2 (Rv1821), Rv3659c
Cytochromes	CcdA (Rv0527), CcsA (Rv0529), CtaB (Rv1451), CtaC (Rv2200c), CtaD (Rv3043c), CtaE (Rv2193), CydA (Rv1623c), CydB (Rv1622c), CydC (Rv1620c), CydD (Rv1621c), Cyp121 (Rv2276), Cyp123 (Rv0766c), Cyp124 (Rv2266), Cyp125 (Rv3545c), Cyp126 (Rv0778), Cyp128 (Rv2268c), Cyp130 (Rv1256c), Cyp132 (Rv1394c), Cyp135A1 (Rv0327c), Cyp135B1 (Rv0568), Cyp136 (Rv3059), Cyp137 (Rv3685c), Cyp138 (Rv0136), Cyp139 (Rv1666c), Cyp140 (Rv1880c), Cyp141 (Rv3121), Cyp142 (Rv3518c), Cyp143 (Rv1785c), Cyp144 (Rv1777), Cyp51 (Rv0764c), DipZ (Rv2874), LldD1 (Rv0694), LldD2 (Rv1872c), QcrB (Rv2196), QcrC (Rv2194), SdhC (Rv3316)

Table 4.3 Prioritised drug targets. The list of candidate drug target proteins of *Mycobacterium tuberculosis H37Rv* sorted by maximum flow values

Candidate protein	Functional class	Pathway name	Max-flow
Rv3457c	information pathways	Transcription	70861
Rv0651	information pathways	(00970) Aminoacyl-tRNA biosynthesis	61319
Rv0732	cell wall and cell processes	Sec-dependent translocation pathway	49050
Rv2476c	intermediary metabolism and respiration	00250 (Alanine, aspartate and glutamate metabolism)	36008
Rv1908c	virulence, detoxification, adaptation	isoniazid-resistance[146]	33297
Rv1837c	intermediary metabolism and respiration	(00630) Glyoxylate and dicarboxylate metabolism	31666
Rv1599	intermediary metabolism and respiration	histidine biosynthesis	30402
Rv2455c	intermediary metabolism and respiration	Citrate cycle (TCA cycle)	27638
Rv1390	information pathways	03020 (RNA polymerase)	27485
Rv3859c	intermediary metabolism and respiration	00250 (Alanine, aspartate and glutamate metabolism)	26942
Rv0002	information pathways	00230 Purine metabolism; 00240 Pyrimidine metabolism; 03030 DNA replication	24226
Rv2391	intermediary metabolism and respiration	Sulfur metabolism	23588
Rv2150c	cell wall and cell processes	00550 (Peptidoglycan biosynthesis)	23472
Rv2534c	information pathways	Translation	20846
Rv1712	intermediary metabolism and respiration	00240 (Pyrimidine metabolism)	20292

Candidate protein	Functional class	Pathway name	Max-flow
Rv2412	information pathways	03010 Ribosome	18318
Rv2847c	intermediary metabolism and respiration	00860 (Porphyrin and chlorophyll metabolism); 01100 (Metabolic pathways)	17115
Rv1415	intermediary metabolism and respiration	riboflavin biosynthesis	16970
Rv3601c	intermediary metabolism and respiration	00770 (Pantothenate and CoA biosynthesis)	16574
Rv3795	cell wall and cell processes	Cell wall biosynthesis	15701
Rv0288	cell wall and cell processes	03070 (Bacterial secretion system)	15573
Rv1303	cell wall and cell processes	Transcriptional Regulation	14635
Rv2151c	cell wall and cell processes	00550 (Peptidoglycan biosynthesis)	14426
Rv3808c	cell wall and cell processes	00550 (Peptidoglycan biosynthesis)	13660
Rv1315	cell wall and cell processes	00550 (Peptidoglycan biosynthesis)	12471
Rv3793	cell wall and cell processes	Lipoarabinomannan biosynthesis	12361
Rv2553c	cell wall and cell processes		12177
Rv0556	cell wall and cell processes		11269
Rv3019c	cell wall and cell processes	Immunomodulation	11899
Rv1602	intermediary metabolism and respiration	histidine biosynthesis	11326
Rv2552c	intermediary metabolism and respiration	01063 (Biosynthesis of alkaloids derived from shikimate pathway)	11190
Rv0511	intermediary metabolism and respiration	00860 (Porphyrin and chlorophyll metabolism)	11111

Candidate protein	Functional class	Pathway name	Max-flow
Rv2868c	conserved hypotheticals	2-C-methylerythritol 4-phosphate (MEP) pathway	10984
Rv0066c	intermediary metabolism and respiration	00020 (Citrate cycle (TCA cycle))	10214
Rv3792	cell wall and cell processes	Arabinan biosynthesis	10093
Rv0051	cell wall and cell processes	cobalamin biosynthesis	9645
Rv2231c	intermediary metabolism and respiration	cobalamin biosynthesis	9603
Rv0292	cell wall and cell processes	03070(Bacterial secretion system)	9588
Rv1547	information pathways	DNA replication	8903
Rv2773c	intermediary metabolism and respiration	00300 Lysine biosynthesis,	8895
Rv2703	information pathways	Transcription Regulation	7637
Rv1738	conserved hypotheticals	[355]	8443
Rv2097c	intermediary metabolism and respiration	03050 (Proteasome)	8323
Rv1711	conserved hypotheticals	pseudouridine synthesis	8320
Rv3490	virulence, detoxification, adaptation	00500 (Starch and sucrose metabolism)	8150
Rv3780	conserved hypotheticals	Secretion system	7976
Rv3794	cell wall and cell processes	Cell wall biosynthesis	7959
Rv0286	PE/PPE	03070 (Bacterial secretion system)	7923
Rv3660c	virulence, detoxification, adaptation	Post-transcriptional regulation	7867
Rv0058	information pathways	DNA replication	6161
Rv2538c	intermediary metabolism and respiration	shikimate pathway	7630

Candidate protein	Functional class	Pathway name	Max-flow
Rv3201c	information pathways	DNA Repair and Replication pathway	7555
Rv1797	cell wall and cell processes	Transport	7467
Rv3240c	cell wall and cell processes	Bacterial secretory system	7412
Rv2093c	cell wall and cell processes	twin-arginine translocase (TAT) secretion pathway	7084
Rv0467	intermediary metabolism and respiration	00630 (Glyoxylate and dicarboxylate metabolism)	7063
Rv1794	conserved hypotheticals		6895
Rv0736	information pathways	03022 (Basal transcription factors)	6830
Rv1044	conserved hypotheticals	Transcriptional regulation	6793
Rv2111c	intermediary metabolism and respiration	03050 (Proteasome)	6593
Rv1382	cell wall and cell processes		6519
Rv0290	cell wall and cell processes	03070(Bacterial secretion system)	6424
Rv0260c	regulatory proteins	mtu00860 Porphyrin and chlorophyll metabolism	6423
Rv1024	cell wall and cell processes	cell division	6359
Rv3805c	cell wall and cell processes	Cell wall biosynthesis	6353
Rv1845c	cell wall and cell processes	Beta-Lactam resistance	6207
Rv0289	cell wall and cell processes	03070(Bacterial secretion system)	6147
Rv0412c	cell wall and cell processes		6069
Rv1159	cell wall and cell processes	PIM Biosynthesis	5942
Rv2191	information pathways	03430 (Mismatch repair)	5389
Rv3531c	conserved hypotheticals	Colesterol metabolism	5228

Candidate protein	Functional class	Pathway name	Max-flow
Rv1575	insertion seqs and phages		5046
Rv0541c	cell wall and cell processes	Protein Mannosylation	5028
Rv0383c	cell wall and cell processes	Protein export	4965
Rv2112c	intermediary metabolism and respiration	03050 (Proteasome)	4608
Rv1783	cell wall and cell processes	ESX-5 secretion system	4580
Rv0249c	intermediary metabolism and respiration	00020 (Citrate cycle (TCA cycle))	4556
Rv3244c	cell wall and cell processes	MtrAB signal transduction pathway	4449
Rv3658c	cell wall and cell processes	03070 (Bacterial secretion system)	4380
Rv0902c	regulatory proteins	(02020)Two component system	4376
Rv2813	conserved hypotheticals	ATP Dependent protein secretion systems	4365
Rv1587c	insertion seqs and phages		4339
Rv0883c	conserved hypotheticals		4335
Rv2748c	cell wall and cell processes	cell division	4305
Rv0164	conserved hypotheticals	03070 (Bacterial secretion system)	4273
Rv0451c	cell wall and cell processes		3278
Rv2647	insertion seqs and phages	Transposition Pathway	4187
Rv0236c	cell wall and cell processes	lipoarabinomannan biosynthesis	4185
Rv0431	cell wall and cell processes		3717
Rv2418c	conserved hypotheticals		3607
Rv2710	information pathways	Transcription Regulation	3597
Rv2492	conserved hypotheticals		3481

Candidate protein	Functional class	Pathway name	Max-flow
Rv2711	regulatory proteins	01053 (Biosynthesis of siderophore group nonribosomal peptides)	3420
Rv0311	conserved hypotheticals		2559
Rv0585c	cell wall and cell processes	Signal transduction	3383
Rv1402	information pathways	Replication	3346
Rv0358	conserved hypotheticals		3343
Rv0227c	cell wall and cell processes		2750
Rv0875c	cell wall and cell processes		2653
Rv3843c	cell wall and cell processes	Transcription	2626
Rv2219	cell wall and cell processes		2584
Rv0666	cell wall and cell processes		2567
Rv2050	conserved hypotheticals		2566
Rv1274	cell wall and cell processes		2561
Rv0817c	cell wall and cell processes		2554
Rv1610	cell wall and cell processes		2550
Rv1343c	cell wall and cell processes		2546
Rv1854c	intermediary metabolism and respiration	00190 (Oxidative phosphorylation)	2540
Rv0955	cell wall and cell processes		2486
Rv3635	cell wall and cell processes		2445
Rv2474c	conserved hypotheticals		1857
Rv2751	conserved hypotheticals		1807
Rv0088	lipid metabolism	polyketide synthesis.	1786
Rv3593	cell wall and cell processes	00312 (beta-Lactam resistance)	1713
Rv3604c	cell wall and cell processes	Transcriptional regulation	1680
Rv2986c	information pathways		1679
Rv2186c	conserved hypotheticals		1675

Candidate protein	Functional class	Pathway name	Max-flow
Rv1128c	insertion seqs and phages	[184]	1673
Rv2416c	virulence, detoxification, adaptation	host immuno modulation	920
Rv2339	cell wall and cell processes		847
Rv3597c	information pathways	Transcriptional regulation	845
Rv2518c	cell wall and cell processes		843
Rv2518c	PE/PPE		831
Rv2444c	information pathways	RNA degradation	831

As it has been discussed in the previous analysis, the proteins in the prioritized list which interact with the host were identified using a dataset retrieved from a computational prediction of Homo Sapiens- *Mycobacterium tuberculosis H37Rv* protein-protein interactions [101]. This was carried out in order to understand the infection mechanism of the pathogen. The resulted 15 out of 131 proteins have been shown in Table 4.4 with their corresponding maximum flow value.

Table 4.4 Potential targets that interact with human

Protein	Functional Class	Max-Flow
Rv0732	cell wall and cell processes	49050
Rv1908c	virulence, detoxification, adaptation	33297
Rv1599	intermediary metabolism and respiration	30402
Rv2455c	intermediary metabolism and respiration	27485
Rv2150c	cell wall and cell processes	23472
Rv2534c	information pathways	20846
Rv1712	intermediary metabolism and respiration	20292
Rv2987c	intermediary metabolism and respiration	17639
Rv1415	intermediary metabolism and respiration	16970
Rv3601c	intermediary metabolism and respiration	16574
Rv2553c	cell wall and cell processes	12177
Rv3921c	cell wall and cell processes	12106
Rv1602	intermediary metabolism and respiration	11326
Rv1611	intermediary metabolism and respiration	10147
Rv2538c	intermediary metabolism and respiration	7630

Candidate proteins at the top 10% maximum flow as lower limit of the 90th percentile which contain 14 proteins (Table 4.5) were taken for further analysis. The targetablity of a drug target depends on several factors which include: essentiality to the growth and survival of the pathogen, non homologous to the host, availability of three-dimensional structure and gene expression. A target would be desirable if it is expressed in the organism at least under disease conditions. The identified potential targets through this analysis are obviously essential, non homologous to human and found with in close neighbourhood of the centre of gravity of the protein-protein interaction network. Out of the top 14 proteins 5 of them have solved three-dimensional structures. Structures of proteins of the pathogen that do not have experimentally solved crystal structures can also be obtained using theoretically calculated structural models.

Table 4.5 Top 14 candidate protein drug targets of *Mycobacterium tuberculosis H37Rv*

Gene Symbol	Locus	Functional Class	Max-Flow	Cross Reference (PDB)
rpoA	Rv3457c	information pathways	70861	
rpsC	Rv0707	information pathways	61418	
rplJ	Rv0651	information pathways	61319	
secY	Rv0732	cell wall and cell processes	49050	
gdh	Rv2476c	intermediary metabolism and respiration	36008	
katG	Rv1908c	virulence, detoxification, adaptation	33297	1SFZ;1SJ2;2CCA;2CCD;4C50;4C51;
glcB	Rv1837c	intermediary metabolism and respiration	31666	2GQ3;3S9I;3S9Z;3SAD;3SAZ;3SB0;
hisD	Rv1599	intermediary metabolism and respiration	30402	
korA	Rv2455c	intermediary metabolism and respiration	27638	
rpoZ	Rv1390	information pathways	27485	
gltB	Rv3859c	intermediary metabolism and respiration	26942	
dnaN	Rv0002	information pathways	24226	3P16;3RB9;
sirA	Rv2391	intermediary metabolism and respiration	23588	1ZJ8;1ZJ9;
ftsZ	Rv2150c	cell wall and cell processes	23472	1RLU;1RQ2;1RQ7;2Q1X;2Q1Y;4KWE;

Literature review has been carried out for these top 14 proteins to find out if the result of this analysis is in line with similar studies. All of the top 14 proteins are Tropical Disease Research (TDR) validated targets for *Mycobacterium tuberculosis H37Rv*. TDR Targets database is a dedicated database to facilitate the rapid identification and prioritization of molecular targets for drug development, focusing on pathogens responsible for neglected human diseases [104]. It has also been observed that some of these proteins were reported as potential drug targets of the pathogen by various studies. An integrated approach of an interactome, reactome and genome-scale structural analysis to identify potential drug targets of *Mycobacterium tuberculosis* is among these studies which implements multi-step filters [65]. The final list of potential drug targets from this study are regarded as reliable targets since the pipeline incorporates a network analysis of the protein-protein interaction, a flux balance analysis of the reactome, experimentally derived phenotype essentiality data, sequence analyses and a structural assessment of targetablity. From our list of top 14 candidate proteins, secY (Rv0732), katG (Rv1908c), gltB (Rv3859c) and sirA (Rv2391) are among the final list of potential targets identified by this study [65]. katG (Rv1908c) is a validated drug target and multifunctional enzyme, exhibiting both a catalase, a broad-spectrum peroxidase, and a peroxynitritase activities. secY (Rv0732) is significantly more useful as a drug target since it has been involved in the emergence of resistance in the interactome by mediating the flow of information from the existing drugs to the resistance machinery. Rv2455c is among the enzymes identified as drug targets of the pathogen by using in silico analysis of Metabolic Pathways [41]. DNA polymerase III β sliding clamp's ability to function with diverse DNA repair proteins and cell cycle-control proteins make it a potential drug target [126]. ftsZ (Rv2150c), a bacterial tubulin homologue involved in essential cell division, is considered as an attractive target to develop novel anti-TB drugs, as well as new broad-spectrum antibacterial agents [127].

The schematic diagram which depicts the concept about the mechanism of resistance to the existing drugs and the way to tackle the problem with the newly proposed targets has been shown in Figure 4.4. Decades-old drugs are no longer effective in killing drug-resistance *mycobacterium tuberculosis* which leads to the requirement of new targets to tackle the problem. The newly proposed targets in this analysis are central to interactome network, essential to the growth and survival of the pathogen. They have been further prioritized based on their maximum flow value to resistance genes. The druggability of proteins with higher maximum flow value is believed to be increased since the inhibition of the target would disrupt the communication.

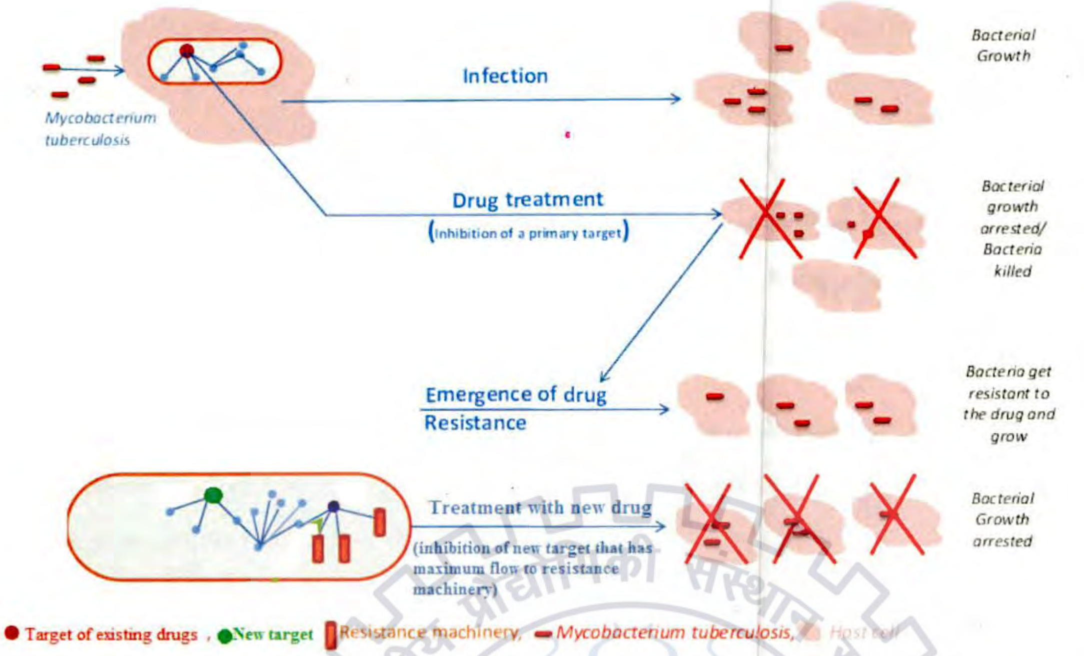

Mycobacterium tuberculosis
Infection
Bacterial Growth
Drug treatment
(Inhibition of a primary target)
Bacterial growth arrested/ Bacteria killed
Emergence of drug Resistance
Bacteria get resistant to the drug and grow
Treatment with new drug
(inhibition of new target that has maximum flow to resistance machinery)
Bacterial Growth arrested
● Target of existing drugs , ●New target ▮Resistance machinery, ▬ Mycobacterium tuberculosis, Host cell

Figure 4.4 Schematic diagram to depict the proposed mechanism to tackle the problem of resistance (adopted from Raman et al. (2008) [44]).

4.4 Assessment of the Method

Validation is one of the challenges in identifying drug targets of TB using computational methods mainly due to the unavailability of standard dataset. It would also be ideal to have negative dataset in order to assess their performances effectively. However, such data is not available because of the lack of interest of researchers in validating them [46]. Yeh et al. (2012) evaluated the performance of maximum flow approach in identifying drug targets of prostate cancer. This has been carried out through a comparison with other approaches: degree, network entropy, betweenness, closeness and random walk [46]. They obtained the highest mean average precision for maximum flow approach which indicates the method has a better performance than the stated approaches.

In this analysis, validated drug targets for currently existing drugs and essential non- homologous proteins were taken to assess the performance of the method. The 8 clinically used drugs for the treatment of tuberculosis interact with 35 different proteins in the proteome network of the pathogen [41]. Out of these proteins, 34 of them were found in the generated proteome network and they were taken as validated drug targets. From the comparative analysis, 572 proteins were identified. Out of them 537 found in the generated network and taken as essential and non-homologous proteins. Then, the average ranks of the validated drug targets on the non-homologous essential proteins were computed based on their maximum flow value to resistance genes. A lower average rank indicates a better performance. The absolute count of validated drug targets in the top 1% of non-homologous essential proteins (practically in the top 5 proteins), in the top 5% (practically in the top 26 proteins), in the top 10% (practically in the top 53 proteins), in the top 15% (practically in the top 80 proteins) and in the top 20% (practically in the top 107 proteins) were reported (Table 4.6). For instance, in the top 1%, the method identified 2 validated drug targets.

Table 4.6 Number of drug targets in top 1%, 5%, 10%, 15% and 20% of the essential non-homologous proteins

Non-homologous Essential proteins	Number of Validated Drug Targets
1%(5)	2
5%(26)	4
10%(53)	13
15%(80)	21
20%(107)	21

4.5 Conclusion

In this study, potential drug targets and resistance genes to the existing drugs of *Mycobacterium tuberculosis H37Rv* were taken as an input and maximum flow approach has been used to prioritize these proteins based on the flow value of each protein to resistance genes. The potential drug target proteins taken as an input are essential to the survival and growth of the pathogen, non-homologous to human and found near to the center of gravity of interactome network. Resistance genes are curated list of reported genes which are involved in both intrinsic and extrinsic drug resistance mechanisms of *Mycobacterium tuberculosis H37Rv* [44]. Using maximum flow approach as a new method on protein-protein interaction network of the pathogen has an importance of considering the flow instead of shortest paths like other global network centrality measures. There are many well established criteria for assessing the targetablity of potential drug target proteins of various diseases which include essentiality, non homologous to the host, availability of solved structure and gene expression under disease conditions. However to our knowledge there is no comprehensive effort to incorporate the influence of drug targets on resistance genes of diseases like *Mycobacterium tuberculosis H37Rv* as criteria of druggablity. This leads to the possibility of including the influences of newly proposed targets on resistance genes as a new concept to assess the druggability of a target. We successfully prioritised potential drug targets based on their flow to resistance genes of the existing drugs of the pathogen. This is believed to increases the success rate of the potential drug targets in the rational drug discovery process. A detail literature review of the top 14 drug targets has also showed that many of these proteins have been suggested as potential drug targets.

CHAPTER FIVE

Identifying potential co-targets of *Mycobacterium tuberculosis H37Rv* using maximum flow approach

5.1 Introduction

In the past, it has been tried to tackle the problem of drug-resistance mainly through various strategies involving wet-lab experiments and clinical decisions [43]. There are only very few success stories that have been achieved through the wide implementation of these strategies. There is a worldwide rise of resistance forms and TB remains one of the main human health threat by being the most widely spread infectious disease. The ineffectiveness of previously implemented strategies highlights the requirement of new rational approaches to tackle the problem from its root cause. Identifying new and effective method to deal with the problem of resistance is more difficult due to our limited knowledge about the resistance triggering mechanisms [44]. The availability of molecular interactome networks for various organisms is an opportunity to explore more about the biological mechanisms at the cellular level. The mechanisms of interactions among genes and their products at the system level can be obtained from the interactome networks. Through the analyses of molecular-level interaction networks, we can enhance our understanding about the complexity of biological systems and be able to reveal hidden relationships among drugs, genes, proteins, and diseases. It can be helpful in explaining specific biological actions like resistance triggering mechanisms. It is desirable to first understand the possible causes and mechanisms of resistance, and then design a rational way to prevent its emergence.

There are few comprehensive computational investigations focused on the emergence of resistance mechanisms of TB and the possible counter measures [44, 68, 69]. Raman and Chandra (2008), tried to identify possible pathways that may be responsible for generating drug resistance using proteome network of *Mycobacterium tuberculosis H37Rv* [44]. They identified controlling hubs within these paths and proposed potential co-target proteins to counter the problem of resistance. Chen et al (2012), uses a random walk model on interactome network of *mycobacterium tuberculosis* and gene expression data to identify potential co-targets of isoniazid and ethionamide [69]. These methods did not consider the non-shortest paths that could be important in the communication of information between drugs and resistance genes.

In the previous analysis (chapter 4), an attempt has been made to prioritize potential drug targets based on their influence on resistance genes to tackle the problem of drug-resistance in a more rational way. This analysis has been carried out to explore possible routes of drug resistance in *Mycobacterium tuberculosis H37Rv* through which information required for triggering drug resistance may be passed on in the cell and to identify proteins that are highly involved in its emergence by mediating information among drug target proteins and resistance genes. In the current treatment system of tuberculosis, there are eight clinically used drugs such as amikacine, isoniazid, ethambutol, pyrazinamide, rifampin, streptomycin, ethionamide and ofloxacin [68]. Proteins involved in the emergence of drug-resistance of these drugs have been proposed as potential co-targets. They can be inhibited in combinations with primary targets simultaneously to disrupt communications to resistance genes in the protein-protein interaction network. It has been implemented with maximum flow approach. Drug-specific weighted protein-protein interaction networks of the pathogen have been used as interactome network dataset where drug targets of each drug have been considered as source nodes, the curated list of genes involved in intrinsic and extrinsic drug resistance mechanisms as sink nodes. The maximum flow values of all proteins other than the drug targets and resistance genes have been computed in the maximum flow from source nodes to sink nodes of drug specific protein-protein interaction networks. Then, proteins were prioritized based on their maximum flow value to the sink node of each drug specific network. A subsequent filters including non-homologous assessment to avoid host toxicity, identification of proteins that interact with the host and essentiality analysis were carried out on the resulted lists of proteins. The refined lists of proteins which have strong influence on the resistance genes were proposed as potential co-targets for each anti-tuberculosis drug. A further betweenness centrality measure analysis has been done on the resulted potential co-targets of isoniazid as a case study by aiming to obtain more reliable co-targets. From the analysis, it has been observed that the identified potential co-target proteins have a strong involvement in the emergence of drug resistance by being used as a mediator of information from drug targets to resistance machineries. So, targeting them with a systematic combination of the existing drugs is believed to be effective to prevent the emergence of drug resistance.

5.2 Materials and Method

Protein-protein interaction network refers to the assembly of the protein signal cascades in which the biological function and information are transferred [128]. Proteins entirely operate with other molecules such as low molecular weight compounds, lipids, nucleic acids, or other proteins on the basis of these interactions [70]. They rarely act alone where the close association with partner proteins or assembly into larger protein complexes is necessary for biological activity. In order to construct drug-specific weighted protein-protein interaction networks of *Mycobacterium tuberculosis H37Rv,* the interaction dataset retrieved from Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) database have been used [96]. Only a portion of the dataset with combined score value of pair of interactions greater than 770 has been considered since it is stated to be more reliable [110]. The combined score values of the interacting pairs of proteins have been assigned as the weight of the respective edges to quantify the possible flow [124].

Drug targets and resistance gene sets have been taken as inputs. Drug targets are proteins related with the eight first line and second line drugs used in the current treatment of TB. Since metabolic adjustments often occur to minimize the effect of inhibition on the particular protein, multiple proteins in the drug related functional mechanism may also be targeted [129,130]. It is reasonable to consider proteins involved in the whole pathway as the drug targets rather than an individual protein [44]. All these proteins that interact with the current clinically used drugs in the treatment of TB have been considered as drug targets [68]. The list of drug resistance genes is a curated list of genes involved in both intrinsic and extrinsic drug resistance mechanisms of *Mycobacterium tuberculosis H37Rv*. A list consisting of 82 genes that are involved in these drug resistance mechanisms has been identified from two published papers [44, 68].

After identifying drug targets of the current clinically used drugs and resistance genes of the pathogen, weighted drug-specific protein-protein interaction networks are then constructed. The drug targets are connected to artificially created dummy source nodes with the capacity of corresponding edges set to infinity. All resistance genes are connected to artificially created dummy sink nodes with the capacity of these edges set to infinity. The weight of the edges connected to source and sink nodes are set to a number that exceeds the potential flow value which is the product of maximum degree and maximum capacity plus one instead of setting them to infinity for the sake of practicality.

The main objective of this analysis is to identify proteins that are involved in mediating information between drug targets of currently used drugs and resistance genes which lead to the development of drug-resistance. The systematic combinations of resulted proteins with primary targets are believed to be useful to prevent the emergence of drug-resistance. This will allow the main drug to kill the bacteria effectively. Maximum flow approach has been used as a main method. The maximum flow values of all proteins in the flow from drug targets to resistance genes were computed in each drug-specific protein-protein interaction network. Then, proteins have been sorted based on these values and those which found at the top were proposed as co-targets of the drug under consideration. The approach was effectively used by Yeh et al. (2012) to predict drug targets of prostate cancer from microarray data, disease genes and interactome network [46]. It has also been used in our previous analysis (chapter 4) to further prioritize potential drug targets of *Mycobacterium tuberculosis H37Rv* [131].

The classical combinatorial maximum flow problem has various methods of implementation; network simplex method of Dantzig [82,83], the augmenting path method of Ford and Fulkerson [84], the blocking flow method of Dinitz [85], and the push–relabel method of Goldberg and Tarjan [86,87]. We have used a modified version of push–relabel method of Goldberg and Tarjan since it has practical superior computational time [81]. The modifications were applied to convert the problem into classical maximum flow problem [46]. In the previous chapter, these modifications have been discussed in detail. The approach was used to further prioritize already identified potential targets of *Mycobacterium tuberculosis H37Rv* based on their influence to resistance genes. The inhibition of a protein with higher maximum flow value was expected to disrupt the communication with resistance genes. Therefore, it could be taken first as a drug target in the expensive and complicated drug discovery process. In this analysis, the maximum flow approach has been used to identify potential co-targets that can be paired with the existing first line and second line drugs of TB for the possible prevention of the emergence of drug resistance. The maximum flow values were computed for all proteins other than the drug target proteins and resistance genes instead of only for drug targets in the possible flow between drug targets and resistance genes. This was done by aiming to identify those proteins that are highly involved in mediating information between drugs and resistance genes so that this bridge of communication can be systematically blocked.

After prioritized list of proteins based on their maximum flow in the flow from respective drug targets to resistance genes in each drug-specific weighted protein-protein interaction network, subsequent assessments have been carried out. The lists have been filtered out through BLASTp search against the non-redundant database with an e-value threshold cut off set to 0.005 by restricting the search to H. sapiens [88]. This has been done to identify proteins that have no detectable homology with human so as to prevent host toxicity. BLAST search against DEG has also been carried out on the resulted lists of non-homologous proteins to identify essential genes. The final prioritized lists of proteins have been proposed as potential co-targets for eight clinically used drugs of TB that can be used with primary targets to prevent the emergence of drug resistance. A further analysis of betweenness centrality measure has been carried out on the proposed co-targets list of isoniazid to numerically characterize the importance of proteins in the biological system where proteins that are more central in the network were expected to be distinguished.

5.3 Results and Discussion

Various forms of resistance to the existing drugs is an overwhelming problem in the treatment of TB. Mutation is one of the main mechanisms of action for adaptation by the bacteria in response to drugs [44]. The bacteria do this to minimize the effect of drugs by reducing their bio-availability and binding to mycolic acid pathway. There has to be information flow from the drug target(s) of a given drug to the molecular components of the resistance machinery in the form of structural and biochemical signals that lead to eventual development of resistance. It is vital to identify the possible routes through which information required for triggering drug-resistance flows in the cell. This will provide an insight about the main mediators for the flow of information. These intermediate proteins which are highly involved in these routes were identified and proposed as potential co-targets so that they can be inhibited in combination with the corresponding primary targets to prevent the emergence of drug resistance.

This analysis has been carried out to identify the stated plausible routes of information from eight clinically used drugs in the current treatment of TB to resistance machineries in each drug-specific protein-protein interaction network of *Mycobacterium tuberculosis H37Rv*. Proteins that have found to be key mediators of information in these pathways have been proposed as potential co-targets for each drug. Accordingly, drug-specific weighted protein-protein interaction networks have been constructed. Maximum flow approach has been used on the generated proteome network to identify potential co-targets. The resulted potential co-targets can be used intelligently to design roadblocks to prevent the emergence of drug resistance.

5.3.1 Weighted Protein-Protein Interaction Network

The protein-protein interaction network dataset of *Mycobacterium tuberculosis H37Rv* was retrieved from STRING [96]. The interaction network consists of both structural and functional linkages among various protein molecules, including indirect linkages. The portion of the dataset with only a combined score value greater than 770 has been considered since it has been shown that it is more reliable by having correlated gene expression profiles and coherent informative GO term annotations in both interaction partners [110]. Some adjustments have been made on the resulted network dataset to derive the required drug-specific weighted protein-protein interaction networks. The combined score values have been used as weights of edges between interacting proteins. Dummy source nodes that connect drug targets of each drug and dummy sink nodes that links resistance genes together were artificially created. The eight clinically used drugs interact with 35 different proteins due to an often occurrence of metabolic adjustments to minimize the effect of inhibition on the particular protein [68, 69]. The targets of each drug and the interacting proteins were taken as initial target sets. Out of these proteins, 34 of them were found in the refined network. Then, these proteins were connected to artificially created dummy source nodes with an associated edge weight values of maximum possible value. The comprehensive sets of targets taken for this analysis have been shown in Table 5.1. In the table, proteins highlighted with bold font are targets where as others are those interacting with each drugs. The dummy sink nodes of the drug specific networks were created by connecting the curated list of resistance genes. A curated list consisting of 82 resistance genes were obtained from references [44] and [68]. Out of them, 78 were found in the constructed protein-protein interaction network, which were considered in this analysis. Whenever there are overlapping genes between drug target proteins and curated list of resistance genes in each drug specific network, they have only been connected to dummy source node.

Table 5.1 List of drug targets and interacting proteins of each of individual drugs

Drugs	Targets
Amikacine	**Rv1694(tlyA)**
Isoniazid	Rv0340, Rv0341, Rv0342, Rv0343, Rv1483, **Rv1484(InhA)**, Rv1592c, Rv1772, Rv1854c, **Rv1908c(KatG)**, Rv1909c, Rv2243, Rv2245, Rv2247, Rv2428, Rv2846c, Rv3139, Rv3566c, Rv3795
Ethambutol	Rv0341, Rv0342, Rv0343, Rv1267c, Rv3124, Rv3264c, Rv3266c, Rv3793, Rv3794, **Rv3795(EmbB)**
Pyrazinamide	**Rv2043c (PncA)**
Rifampin	**Rv0667(RpoB)**, Rv2629, Rv3795
Streptomycin	**Rv0682(RpsL)**, Rv3919c
Ethionamide	Rv3854c, **Rv1484(InhA)**
Ofloxacin	**Rv0005(GyrB), Rv0006(GyrA)**

Through the stated adjustments, eight drug specific protein-protein interaction networks were constructed. These networks were used to identify potential co-targets of the eight clinically used first line and second line drugs. As a case, the general statistical properties of isoniazid-specific network have been discussed here. The network contains 14,766 undirected interactions among 3,487 proteins. The statistical properties of the network have been shown in Table 5.2.

Table 5.2 Statistical properties of isoniazid-specific protein-protein interaction network

Parameter	Value
Number of nodes (n)	3487
Connected components	106
Network diameter	15
Average number of neighbours	8.469
Network density	0.002
Characteristic path length	5.575
Clustering coefficient	0.379

Characteristic path length of the network, an average of shortest paths between interacting pairs of proteins, is small in spite of the network's size. It shows that the network is in compact form where the communication between any two vertices is possible through traversing only few steps. It is also smaller than $log(n)$, where *n* is the network's size, which implies that the generated *Mycobacterium tuberculosis H37Rv* proteome interaction network is a small world network. The shortest path length distribution of the network is shown in Figure 5.1. The network's *clustering coefficient* which shows local cohesiveness and internal structure of the given network is significantly higher than the *clustering coefficient* of a random graph with the same number of vertices. Another important characteristic of a network is its degree distribution *p* (*k*) which is the measure of the proportion of nodes in the network having degree k. The generated isoniazid-specific protein-protein interaction network is a scale free network since its degree distribution approximately follows a power law $p\ (k) = k^{-\gamma}$, with γ~1.758. The corresponding degree distributions are shown in Figure 5.2.

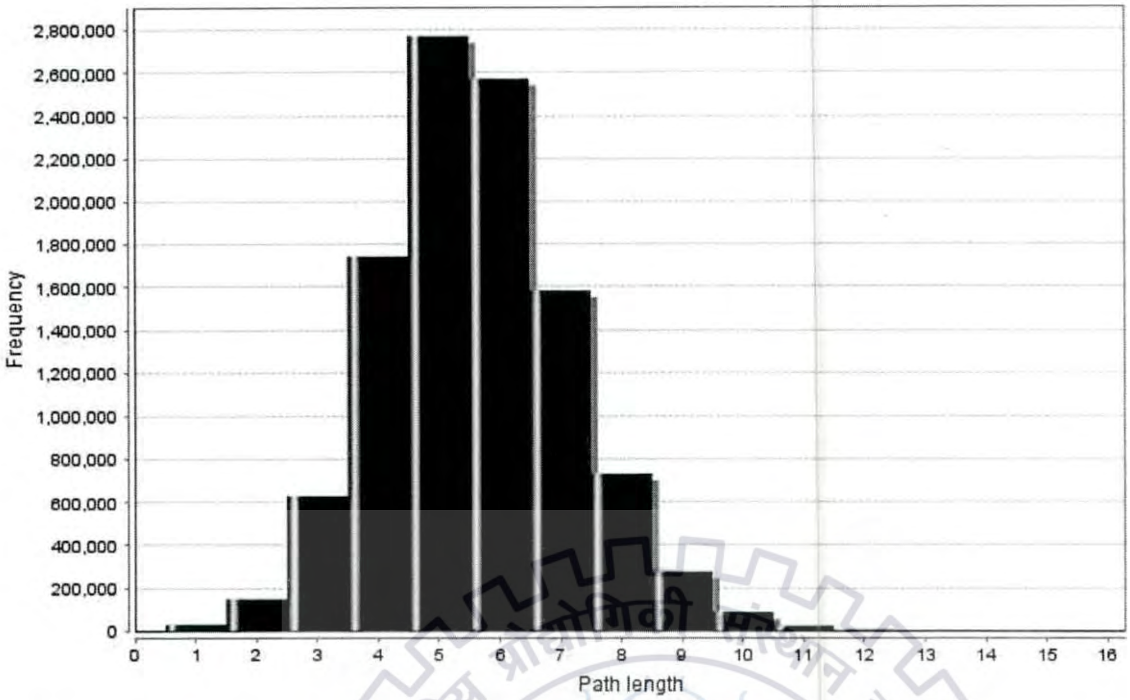


Figure 5.1 Shortest path length distributions. The diagram shows the distribution of shortest path lengths of pair-wise protein interactions.

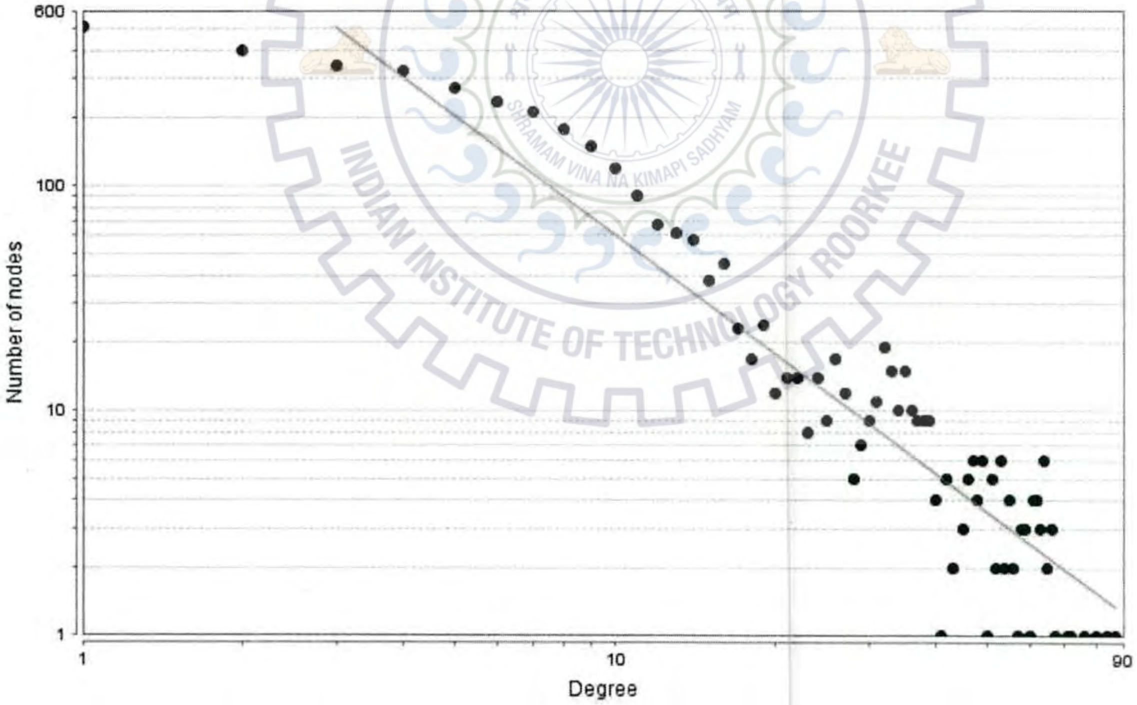


Figure 5.2 Node degree distributions. The distribution of the probability $p\ (k)$ follows a power law $p\ (k) = k^{-\gamma}$.

5.3.2 Identified co-targets for each drug

The maximum flow values of each protein in the flow from drug targets (source nodes) to resistance genes (sink nodes) were computed in each of the drug-specific networks. Then, the proteins have been sorted based on their maximum flow values. A protein with a higher maximum flow value is hypothesized to be the key in the development of drug resistance. It is the main mediator of communication between drug targets and resistance genes. These proteins are believed to be potential candidates of co-targets for the existing clinically used drugs. The co-targets are aimed to be used in a combination with the primary targets to disrupt the network and reduce the emergence of drug resistance.

The flow of drug targets to sink nodes has been observed by aiming to identify those targets that are more involved in the emergence of drug resistance. The result showed that there is a significance difference in their maximum flow values which range from 0 to 40863. This indicates that some of the targets are more responsible in the emergence of resistance than others. Even though it requires further validation, drug targets that have no flow to resistance genes imply that they are not involved in the development of drug-resistance.

The necessary condition for a given protein to be co-target is having a strong influence in the emergence of resistance by mediating the information flow from the given primary target to resistance genes. However, the protein could have additional features like; it might be reported as a primary target when considered individually, it could be non-homologous to human, it could be essential for the survival and growth of the pathogen, and it could also interact with the host. These features can increase or decrease the targetablity of the potential candidate co-target proteins. Therefore, inhibitions of those proteins that have better property of targetablity are expected to have high probability of becoming a part of successful new anti-tuberculosis therapeutics [132].

As it has been stated, in addition to the involvement of proteins in the development of resistance as main criteria for identification of potential co-targets of clinically used drugs, it is important to consider the possible side effects which will be caused by these co-targets. One approach to this is through modifications of the drug molecule as it has been a common practice in traditional drug discovery [44]. However, dealing with this problem at the target identification phase looks like more effective. So, proteins that have been prioritized using their maximum flow value to resistance genes were blasted against the non-redundant database with an e-value threshold cut off set to 0.005 [88]. The search was also restricted to H. sapiens to identify those proteins which have no detectable homologous with human. Essentiality analysis has been carried out on those proteins which have no detectable sequence level homology with human. Essential genes are genes required to sustain the cellular life of the bacteria in which an inhibition of them will have a high probability of success in killing it. This makes essential genes as commonly accepted potential drug targets for parasitic diseases like TB. BLAST search of protein coding genes for the prioritised non-homologous list of potential co-targets were computed against Database of Essential Genes (DEG) to identify those proteins which are essential for the survival and growth of *Mycobacterium tuberculosis H37Rv* [113-115].

One of the main critics with the protein-protein interaction network dataset is that it is integration of interactions activated under various conditions and it doesn't contain information about the conditions under which the interactions may take place [69]. Which means it is not a real snapshot of the interactions in vivo especially during infection. Identifying a list of proteins of *Mycobacterium tuberculosis H37Rv* that interact with human is essential for understanding the infection mechanism of the formidable pathogen. In two recent classical computational works [101, 102], a few *Mycobacterium tuberculosis H37Rv* gene list which interact with human have been identified. These two gene lists are the current standard datasets of potential drug targets for H37Rv. In our analysis, the proposed co-target lists of proteins have been compared with these datasets. There are only few overlaps since these datasets are not comprehensive.

The resulted lists of potential co-targets for the eight clinically used drugs are large. Hence, the top 50 potential co-targets of each drug have been displayed in the following consecutive tables (Table 5.3 to Table 5.10). The tables contain detailed information about potential co-target proteins of each drug such as maximum flow values, essentiality and interaction with the host. The proteins in the lists are also non-homologous to human. A row with bold font indicates that the protein has interaction with the host.

Table 5.3 Top 50 potential co-targets of isoniazid

Rv number	Max-flow	Essentiality
Rv3316	8639	No-essential
Rv0111	6423	No-essential
Rv0033	5682	No-essential
Rv0133	5619	No-essential
Rv3552	5333	No-essential
Rv1547	5119	Essential
Rv2043c	4439	No-essential
Rv3154	4108	No-essential

Rv3240c	4099	Essential
Rv0533c	3969	No-essential
Rv2938	3825	No-essential
Rv2391	3607	Essential
Rv0732	**3599**	**Essential**
Rv3034c	3529	Essential
Rv3660c	3500	Essential
Rv3474	3458	No-essential
Rv2456c	3326	No-essential
Rv1021	**3279**	**No-essential**
Rv0527	3258	Essential
Rv1599	3097	Essential
Rv0081	2908	No-essential
Rv3321c	2902	No-essential
Rv0710	2797	Essential
Rv0707	2763	Essential
Rv1298	2743	Essential
Rv0517	2728	No-essential
Rv2289	2700	No-essential
Rv3197A	2693	No-essential
Rv0639	**2663**	**Essential**
Rv0016c	2654	No-essential
Rv0169	2643	No-essential
Rv2534c	**2625**	**Essential**
Rv1526c	2615	No-essential
Rv3462c	2612	Essential
Rv0709	2606	Essential
Rv0371c	2604	No-essential
Rv0262c	2603	No-essential
Rv0228	2596	Essential
Rv2081c	2574	No-essential
Rv1334	2554	No-essential
Rv3374	2550	No-essential
Rv0177	2549	No-essential
Rv0368c	2531	No-essential
Rv2351c	2521	No-essential
Rv1769	2519	No-essential
Rv1337	2496	No-essential
Rv2145c	2484	Essential
Rv1765c	2468	No-essential
Rv0372c	2442	Essential
Rv2880c	2438	No-essential

Table 5.4 Top 50 potential co-targets of ethambutol

Rv number	Max-flow	Essentiality
Rv3808c	5541	Essential
Rv3792	4249	Essential
Rv3809c	3721	Essential
Rv3316	3437	Non-essential
Rv2673	3368	Essential
Rv3780	2965	Essential
Rv3806c	2943	Essential
Rv3660c	2864	Essential
Rv2081c	2863	Non-essential
Rv1526c	2786	Non-essential
Rv3265c	2786	Essential
Rv3782	2765	Essential
Rv3805c	2709	Essential
Rv3779	2585	Non-essential
Rv0262c	2582	Non-essential
Rv0018c	2552	Essential
Rv2938	2529	Non-essential
Rv2456c	2523	Non-essential
Rv3817	2508	Non-essential
Rv1973	2500	Non-essential
Rv0557	2407	Essential
Rv2174	2389	Essential
Rv1837c	2386	Essential
Rv2351c	2379	Non-essential
Rv2181	2372	Non-essential
Rv0589	2362	Non-essential
Rv2476c	2349	Essential
Rv0561c	2330	Non-essential
Rv2264c	2320	Non-essential
Rv1867	2303	Non-essential
Rv2843	2297	Non-essential
Rv2939	2297	Non-essential
Rv0541c	2237	Essential
Rv3862c	2228	Non-essential
Rv0051	2225	Essential
Rv1303	2175	Essential
Rv1021	**2161**	**Non-essential**
Rv0515	2104	Non-essential
Rv1599	**2043**	**Essential**
Rv0033	2036	Non-essential
Rv0849	2019	Non-essential
Rv3450c	1971	Non-essential
Rv0008c	1939	Non-essential

Rv3474	1938	Non-essential
Rv0034	1936	Non-essential
Rv1382	1901	Essential
Rv2029c	1898	Non-essential
Rv1680	1887	Non-essential
Rv3034c	1882	Essential
Rv3789	1831	Essential

Table 5.5 Top 50 potential co-targets of amikacine

Rv number	Max-flow	Essentiality
Rv3227	1584	Essential
Rv1885c	1568	No-essential
Rv2537c	1530	Essential
Rv2123	1518	No-essential
Rv2062c	1464	No-essential
Rv2488c	1426	No-essential
Rv2162c	1399	No-essential
Rv3729	1331	No-essential
Rv1881c	1323	No-essential
Rv3519	1315	No-essential
Rv2417c	1204	No-essential
Rv1082	1079	No-essential
Rv0176	1072	No-essential
Rv1733c	1066	No-essential
Rv0515	1056	No-essential
Rv0173	1048	No-essential
Rv2199c	1030	No-essential
Rv1604	1020	No-essential
Rv1691	953	No-essential
Rv1693	938	No-essential
Rv0278c	933	Essential
Rv0793	933	No-essential
Rv1021	919	No-essential
Rv2801c	912	No-essential
Rv1247c	910	No-essential
Rv3357	910	No-essential
Rv3497c	903	No-essential
Rv2302	902	No-essential
Rv0870c	901	No-essential
Rv0499	889	No-essential
Rv2877c	889	No-essential
Rv2880c	879	No-essential
Rv2781c	878	No-essential
Rv1303	876	Essential
Rv2818c	872	No-essential

Rv3494c	872	No-essential
Rv3496c	872	No-essential
Rv3498c	872	No-essential
Rv3500c	872	No-essential
Rv3501c	872	No-essential
Rv3606c	872	No-essential
Rv2819c	868	Essential
Rv2840c	868	No-essential
Rv2146c	867	No-essential
Rv0228	866	Essential
Rv2972c	862	No-essential
Rv2975c	862	No-essential
Rv0002	860	Essential
Rv2412	860	Essential
Rv3103c	857	No-essential

Table 5.6 Top 50 potential co-targets of pyrazinamide

Rv number	Max-flow	Essentiality
Rv0793	2019	Non-essential
Rv2585c	1672	Non-essential
Rv2528c	1653	Non-essential
Rv1556	1629	Non-essential
Rv0203	1619	Non-essential
Rv3369	1614	Non-essential
Rv3240c	1613	Essential
Rv0979A	1611	Non-essential
Rv1254	1610	Essential
Rv0826	1608	Non-essential
Rv1571	1581	Non-essential
Rv1372	1578	Non-essential
Rv2335	1577	Non-essential
Rv0034	1575	Non-essential
Rv2540c	**1566**	**Essential**
Rv1563c	1562	Essential
Rv3423c	1560	Essential
Rv1028A	1559	Non-essential
Rv3834c	**1556**	**Essential**
Rv2307B	1551	Non-essential
Rv3662c	1551	Non-essential
Rv0849	1550	Non-essential
Rv3531c	1548	Essential
Rv0140	1545	Non-essential
Rv1329c	1545	Non-essential
Rv2413c	1521	Essential
Rv0407	1511	Non-essential

Rv1331	1511	Non-essential
Rv1955	1505	Non-essential
Rv1282c	1500	Non-essential
Rv3804c	1500	Essential
Rv2991	1459	Non-essential
Rv0792c	1458	Non-essential
Rv3872	1444	Non-essential
Rv2612c	1387	Essential
Rv1946c	1368	Non-essential
Rv3263	1368	Non-essential
Rv3911	1314	Non-essential
Rv2227	1272	Non-essential
Rv1502	1230	Essential
Rv2954c	1220	Non-essential
Rv2955c	1214	Non-essential
Rv1817	1212	Non-essential
Rv3278c	1203	Non-essential
Rv3844	1195	Non-essential
Rv0890c	1167	Non-essential
Rv2949c	1103	Essential
Rv0018c	1097	Essential
Rv2042c	1095	Non-essential
Rv0022c	1094	Non-essential

Table 5.7 Top 50 potential co-targets of rifampin

Rv number	Max-flow	Essentiality
Rv3474	3368	Non-essential
Rv1254	3003	Essential
Rv1298	2784	Essential
Rv1599	**2671**	**Essential**
Rv0589	2625	Non-essential
Rv1908c	**2623**	**Essential**
Rv0475	2607	Non-essential
Rv2043c	2592	Non-essential
Rv0826	2558	Non-essential
Rv0002	2495	Essential
Rv1547	2469	Essential
Rv2307B	2447	Non-essential
Rv3462c	2432	Essential
Rv2991	2400	Non-essential
Rv3240c	2357	Essential
Rv2159c	2334	Non-essential
Rv1885c	2296	Non-essential
Rv2801c	2171	Non-essential
Rv3406	2160	Non-essential

Rv2626c	2113	Non-essential
Rv2456c	2043	Non-essential
Rv0376c	2031	Essential
Rv2223c	2018	Non-essential
Rv2840c	1994	Non-essential
Rv2029c	1980	Non-essential
Rv2720	1978	Essential
Rv3794	1959	Essential
Rv0517	1949	Non-essential
Rv1299	**1926**	**Essential**
Rv2349c	1919	Non-essential
Rv2351c	1876	Non-essential
Rv1015c	1872	Essential
Rv3181c	1871	Non-essential
Rv2612c	1869	Essential
Rv3793	1856	Essential
Rv3370c	1854	Non-essential
Rv3789	1845	Essential
Rv0707	1808	Essential
Rv1986	1807	Non-essential
Rv2703	1800	Essential
Rv0789c	1789	Non-essential
Rv0216	1756	Essential
Rv3069	1739	Non-essential
Rv1954c	1732	Non-essential
Rv3910	1725	Essential
Rv1758	1721	Non-essential
Rv0172	1706	Non-essential
Rv1502	1701	Essential
Rv3911	1701	Non-essential
Rv3630	1697	Non-essential

Table 5.8 Top 50 potential co-targets of streptomycin

Rv number	Max-flow	Essentiality
Rv1298	5112	Essential
Rv2462c	3626	Non-essential
Rv2720	3445	Essential
Rv1015c	3382	Essential
Rv1254	2946	Essential
Rv0105c	2694	Non-essential
Rv0054	2665	Non-essential
Rv2880c	2582	Non-essential
Rv3462c	2469	Essential
Rv1372	2432	Non-essential
Rv1021	**2405**	**Non-essential**

Rv1110	2291	Essential
Rv0002	2174	Essential
Rv1547	2098	Essential
Rv3316	1970	Non-essential
Rv3921c	**1949**	**Essential**
Rv2800	1895	Non-essential
Rv2840c	1869	Non-essential
Rv0707	1838	Essential
Rv1694	1833	Non-essential
Rv3461c	1832	Non-essential
Rv1867	1829	Non-essential
Rv2412	1820	Essential
Rv2235	1788	Essential
Rv1680	1775	Non-essential
Rv1708	1759	Essential
Rv0651	1758	Essential
Rv0710	1751	Essential
Rv0706	1745	Essential
Rv3393	1728	Non-essential
Rv2545	1723	Non-essential
Rv0475	1722	Non-essential
Rv0722	1717	Essential
Rv3069	1707	Non-essential
Rv1460	1703	Non-essential
Rv0058	1698	Essential
Rv0617	1698	Non-essential
Rv0114	1697	Non-essential
Rv2761c	1691	Non-essential
Rv1805c	1685	Non-essential
Rv0732	**1680**	**Essential**
Rv2785c	1680	Non-essential
Rv2162c	1679	Non-essential
Rv0065	1674	Non-essential
Rv0081	1674	Non-essential
Rv3320c	1674	Non-essential
Rv0637	1673	Essential
Rv1335	1672	Non-essential
Rv1338	1671	Essential
Rv0218	1667	Non-essential

Table 5.9 Top 50 potential co-targets of ethionamide

Rv number	Max-flow	Essentiality
Rv3552	1731	Non-essential
Rv0533c	1727	Non-essential
Rv2673	1685	Essential
Rv1158c	1610	Non-essential
Rv1556	1601	Non-essential
Rv0228	1587	Essential
Rv3166c	1573	Non-essential
Rv3857c	1558	Non-essential
Rv2728c	1557	Non-essential
Rv0034	1552	Non-essential
Rv3163c	1551	Non-essential
Rv3762c	1550	Non-essential
Rv1513	1473	Non-essential
Rv0033	1436	Non-essential
Rv2801c	1395	Non-essential
Rv3662c	1367	Non-essential
Rv2585c	1350	Non-essential
Rv1282c	1346	Non-essential
Rv3462c	1345	Essential
Rv1294	1326	Essential
Rv1939	1326	Non-essential
Rv1936	1324	Non-essential
Rv2376c	1298	Non-essential
Rv0145	1228	Non-essential
Rv0593	1165	Non-essential
Rv1012	1032	Non-essential
Rv3809c	1014	Essential
Rv0017c	1006	Non-essential
Rv1653	**1006**	**Essential**
Rv0997	1001	Non-essential
Rv1690	996	Non-essential
Rv2302	970	Non-essential
Rv0423c	**968**	**Essential**
Rv1197	957	Non-essential
Rv1198	955	Non-essential
Rv3316	952	Non-essential
Rv2066	937	Non-essential
Rv1801	929	Non-essential
Rv2859c	929	Non-essential
Rv1908c	**928**	**Essential**
Rv2029c	928	Non-essential
Rv0022c	924	Non-essential
Rv2551c	922	Non-essential

Rv0292	918	Essential
Rv3660c	912	Essential
Rv2800	911	Non-essential
Rv0649	899	Non-essential
Rv3465	899	Essential
Rv2537c	897	Essential
Rv1973	894	Non-essential

Table 5.10 Top 50 potential co-targets of ofloxacin

Rv number	Max-flow	Essentiality
Rv0002	2918	Essential
Rv3921c	**2333**	**Essential**
Rv3474	2324	Non-essential
Rv2289	2268	Non-essential
Rv0301	2249	Non-essential
Rv1847	2219	Non-essential
Rv0003	**1939**	**Essential**
Rv0001	**1873**	**Essential**
Rv0710	1760	Essential
Rv2442c	1760	Essential
Rv1908c	**1742**	**Essential**
Rv2163c	**1728**	**Essential**
Rv2043c	1717	Non-essential
Rv2344c	1712	Essential
Rv0467	1705	Essential
Rv2166c	1700	Essential
Rv3795	1678	Essential
Rv3462c	1660	Essential
Rv1132	1655	Non-essential
Rv3249c	1655	Essential
Rv0997	1646	Non-essential
Rv2090	1640	Non-essential
Rv0203	1629	Non-essential
Rv3630	1626	Non-essential
Rv1015c	1618	Essential
Rv1026	1617	Essential
Rv2192c	**1612**	**Essential**
Rv3857c	1612	Non-essential
Rv2938	1607	Non-essential
Rv0218	1605	Non-essential
Rv1053c	1598	Non-essential
Rv0561c	1597	Non-essential
Rv2160A	1596	Non-essential
Rv0556	1593	Essential

Rv3605c	1582	Non-essential
Rv2949c	1581	Essential
Rv1456c	1580	Essential
Rv3082c	1571	Essential
Rv2412	1566	Essential
Rv0659c	1564	Non-essential
Rv0114	1559	Non-essential
Rv1694	1559	Non-essential
Rv3922c	1557	Essential
Rv1991c	1556	Non-essential
Rv1520	1550	Non-essential
Rv0979A	1544	Non-essential
Rv1012	1541	Non-essential
Rv0053	1540	Essential
Rv2307B	1534	Non-essential
Rv0188	1532	Non-essential

5.3.3 Further analysis on the potential co-targets of isoniazid

Isoniazid (INH) is one of the most frequently used drugs in the treatment of TB [69]. It is the first-line medication used worldwide in prevention and treatment of TB. It is also in the World Health Organization's list of essential medicines which contains a list of medications required for fulfilment of the bare minimum of a basic health system [133]. INH adducts (INH-NAD(P)) of the pyridine nucleotide coenzymes are the mycobactericidal agents of INH. These agents are generated in *vivo* after INH activation and they bind to inhibit essential enzymes. Isonicotinic acid hydrazide, the powerful and specific antitubercular effects of isoniazid, was discovered in 1952 and revolutionized the treatment of tuberculosis since then [134-137]. INH has been used singly in prophylaxis or in a multi-drug combination with rifampicin, pyrazinamide and ethambutol for active infections. However, an eventual emergence of resistance has been a big challenge in the effectiveness of the drug for the treatment of the disease.

The main focus of this analysis is to identify and propose potential co-targets of isoniazid through systematic combination of methods by aiming the inhibition of newly proposed targets with combination of isoniazid will have a better success in dealing with drug- resistance. The proteins were sorted based on their maximum flow values in the flow from dummy source node to dummy sink node of isoniazid-specific network. Only proteins that have maximum flow values greater than zero were considered for further analysis. A protein with maximum flow value of zero means there is no flow that passes through these proteins from drug targets to resistance genes, thus it is not involved in the development of resistance. Even though the involvement of proteins in the development of resistance is the main criteria for identification of potential co-targets, comparative genome and betweenness centrality measure were carried out on the resulted list of proteins to increase their druggablity. After computing betweenness centrality measures for the resulted filtered list, proteins found at the center of gravity of interactome network were reported as a final list of potential co-targets of isoniazid. For a protein to be called it found at the centre of gravity of the interactome network of interest, its betweenness measure should be above the total number of shortest paths expected to pass through it, which is 19440.025. The list consisting of 371 proteins have been found at the center of gravity of the protein-protein interaction network. The top 10 proteins from this list have been shown in Table 5.11

Table 5.11 Top 10 potential co-targets of isoniazid resulted from further analysis

Rv number	Max-flow	Betweenness	Essentiality
Rv3316	8639	42475.86	No-essential
Rv0111	6423	66460.61	No-essential
Rv0033	5682	76296.38	No-essential
Rv0133	5619	72768.63	No-essential
Rv3552	5333	47026.63	No-essential
Rv1547	5119	34472.22	Essential
Rv2043c	4439	100070.1	No-essential
Rv3240c	4099	20443.37	Essential
Rv2938	3825	34863.73	No-essential
Rv2391	3607	90561.54	Essential

5.4 Assessment of the method

The limitation of this analysis is that it would be ideal if all of the interactions were included in the constructed protein-protein interaction network but due to the general low quality of the major protein-protein interaction network datasets of the pathogen only reliable portion of it was considered. However, the network looks comprehensive enough since it covers most of the proteins in the generated proteome interactome. Additionally, the weight assigned for each pair of interactions might not represent direct signal flow since it is difficult to obtain such kind of data. But the protein-protein interaction weights were derived from the 'combined score' value of each interaction which was calculated from various supporting pieces of evidence and this is a strong indicator about the existence of interactions between proteins under consideration. More importantly, many of the approaches used in the identifications of drug targets of the pathogen favor shortest paths but this approach is an inclusive of all paths as far as they have been involved in the emergence of drug resistance. Previous assessments showed that maximum flow approach has superior performance than similar methods used in drug target identification [46, 131].

5.5 Conclusion

With the increasing availability of protein-protein interaction network, network based analysis is feasible to address the issue of drug resistance from the system's perspective. The network provides a systems-level view of how genes and their products interact within the cell and can be used to explain the biological actions under specific condition like drug. This is expected to be useful in implementing novel network-based approach by identifying effective combination of drugs to tackle the problem of drug-resistance from systems perspective. In this analysis, proteins of *Mycobacterium tuberculosis H37Rv* have been prioritized based on their involvement in the emergence of drug-resistance using maximum flow approach. Identification of proteins involved in mediating information between drugs and drug-resistance genes is an important finding that can be used in designing a systematic way of preventing emergence of drug-resistance through effective target co-target combinations. Since maximum flow approach is based on the flow, it is not affected by biasness towards shortest paths like the common global network centrality measures. To prevent host toxicity that can be caused by the intended co-targets, only proteins which are non-homologous with human were considered. Additional analyses such as essentiality assessment and identification of proteins that interact with the host have been carried out. Identification of proteins that interact with the host will be helpful in understanding infection mechanisms. Through these assessments, lists of potential co-targets of eight clinically used drugs of *Mycobacterium tuberculosis H37Rv* were proposed. These proteins can be used in a combination with the primary targets for preventing the emergence of drug resistance at the initial phase of drug discovery process.

A further analysis of betweenness centrality measure has been carried out on the potential co-targets of isoniazid. Through betweenness centrality measure, proteins likely to be essential for the functioning of the system by serving as a bridge of communication between several other proteins in the network were filtered out to increase the success of identified co-targets. The final list of proposed proteins is hypothesised as a more reliable list that can be used in combination with isoniazid to prevent the emergence of drug resistance.

CHAPTER SIX

Structural analysis of protein translocase subunit SecY from *Mycobacterium tuberculosis H37Rv*: a potential target for anti-tuberculosis drug discovery

6.1 Introduction

In previous chapters, lists of potential drug targets and co-targets of *Mycobacterium tuberculosis H37Rv* have been identified through the analyses of protein-protein interaction network. One of the challenges in the identification of potential drug targets of TB is validation mainly due to the lack of standard and experimentally validated drug target dataset. In those analyses, it has been tried to validate the methods used to identify the proposed potential drug targets. The resulted lists were also compared with the previously reported targets of the pathogen that were generated through other target identification methods. Moreover, druggablity of a specified target depends on various factors such as non-homologous to the host, essentiality to the survival and growth for the pathogen and the availability of protein's three-dimensional structures.

The availability of structure of a protein is vital for follow up analyses such as protein function, interactions, antigenic behaviour and rational design of proteins with increased stability or novel function in addition to its obvious use in structure-based drug design [138]. The basic principle is that protein structures lead to protein function. The distinctive protein structure is important for the allocation of chemical groups in specified three-dimensional space. Due to this placement, proteins catalyze many chemical reactions, play important role in structural, transport and regulatory functions in organism. Since protein functions are diverse, protein structures are also diverse. Proteins commonly interact with each other and with other molecules. These interactions expanded functional diversity.

Crystal structures of proteins can be identified using either experimental methods such as NMR and X-ray crystallography or computational structure prediction methods like comparative modelling, threading and ab initio methods. Experimental structure determination methods are primary preferences because they generate high-resolution structural information. The structural genomic project, established with the aim of solving the structures of all proteins, has achieved reasonable success in determining structures of many proteins [139]. However, the rate of structural genomics in solving macromolecular structures is slow where a large number of proteins still do not have experimentally solved structures [46]. This is becoming a barrier for exploiting the information present in the rapidly expanding sequence database. Additionally, a protein could be too large for NMR or it could not be crystallized by X-ray diffraction. Even if it is possible to crystallize the structure of the protein, it is still expensive, lengthy and labour intensive [140]. Therefore, computational structure prediction methods are more often in use to obtain structural information of proteins that don't have experimentally identified structures. Computational structure prediction methods can be classified in to two types. The first category include threading and homology modelling where structures of proteins are predicted based on detectable similarity spanning most of the modelled sequence and at least one known structure. The other is de novo or ab initio methods that predicts the structure of a protein based on only sequence.

In this study, *in silco* structural analysis of protein translocase subunit SecY(Rv0732) has been carried out to get a descriptive three-dimensional structure of the protein and to identify its active site. An identification and characterization protein-ligand or protein-protein binding site is a vital step in the process of structure-based drug design [141, 142]. But information about the binding site is not always known at the beginning and even if it is known, it may not be a suitable binding pocket for a ligand or inhibitor.

Protein translocase subunit SecY is an essential protein for protein export that interacts with Sec (Rv3240c) and SECE (Rv0638) to allow the translocation of proteins across the plasma membrane, by forming part of a channel [143, 144]. The predicted protein interaction partners of protein translocase subunit SecY (Rv0732) has been retrieved from Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) database [96]. The resulted interactions have been shown in Figure 6.1. Protein translocase subunit SecY (Rv0732) is the central subunit of the protein translocation channel SecYEG which consists of two halves formed by TMs 1-5 and 6-10. A lateral gate at the front is formed by the two domains. This gate open onto the bi-layer between TMs 2 and 7, and are clamped together by SecE at the back. Both a pore ring composed of hydrophobic SecY residues and a short helix (helix 2A) on the extracellular side of the membrane which forms a plug closes the channel. The plug probably moves laterally to allow the channel open. The ring and the core may move independently. It is one of the highly ranked proteins in the list of potential drug targets proposed in our previous analyses. It is also among very few proteins of the pathogen that interact with the host as it has been shown in two classical computational works [101, 102]. Even though, these datasets of proteins that interact with the host are not compressive, it is still a strong indicator of the protein's involvement in the infection mechanism. As a result, the predicted three dimensional model of SecY is expected to be of great help by providing novel target for structure-based drug design against drug-resistance TB.

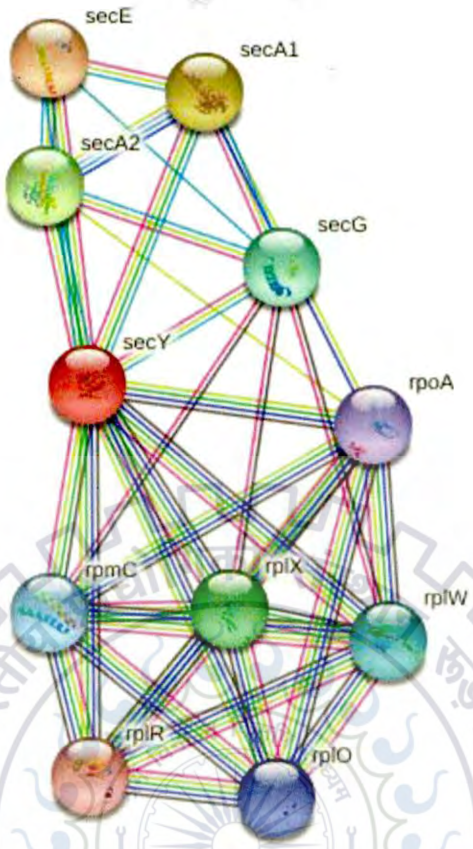


Figure 6.1 Predicted interaction partners of protein translocase subunit SecY [96]

6.2 Methodology

Protein translocase subunit SecY (Rv0732) from *Mycobacterium tuberculosis H37Rv* has been selected from the results of previous analyses on identifying potential targets and co-targets of the pathogen. The protein has been chosen based on predefined criteria which include its rank on the resulted list of potential drug targets and co-targets, unavailability of solved three dimensional structures, non-homologous to the host, essentiality to the growth and survival of the bacteria, verification if it has been reported by other target identification methods and interaction with the host. Then, the primary sequence of the protein was retrieved from TubercuList [144]. TubercuList is genome-derived knowledge-base relational database which serves as a broad source of dataset for analyses *in Mycobacterium tuberculosis* genome. The database is organised from the integration of genome details, drug and transcriptome data, mutant and operon annotation, protein information, bibliography, structural views and comparative genomics. The data is well organised in the way which can be used for rational development of new diagnostic, therapeutic and prophylactic measures against tuberculosis. Subsequently, a multiple sequence alignment was generated among the target sequence and five sequences which are orthologous group members from selected species such as *Arabidopsis thaliana* , *Escherichia coli K12*, *Mycobacterium leprae strain TN, Oryza sativa and Wolbachia endosymbiont of Brugia malayi*. We have used T-Coffee with default parameters to generate the multiple sequence alignments among the sequences [145]. The multiple sequence alignment was performed to observe the differences in pathogenicity among orthologous group members created via successive deletion or insertion of amino acids. Easy Sequencing in PostScript (ESPript) has been used for the graphical enhanced visualisation of the aligned sequences [146]. The phylogenic tree has been inferred from the resulted multiple sequence alignment for the purpose of visualization of changes that has occurred in the evolution of species.

For modelling, we used restrained-based modelling implemented in the program MODELLER 9.11 [91]. This program is an automated approach to comparative modelling by satisfaction of spatial restraints. It uses an alignment of a sequence to be modeled with known related structures to calculate a model containing all non-hydrogen atoms. The modelling process has been started by searching for a template with identified and known three-dimensional structure. The sequences of identified template are then aligned with sequences of the target to identify the conserved regions. Usually, the resulted alignment is the input to the program which predicts the model. The output is a three-dimensional model for the target sequence containing all main-chain and side-chain non-hydrogen atoms. Ramachandran diagram of the resulted model has been generated to visualise the percentage of residues that lie on preferred regions, allowed regions and outliers.

The active site of the predicted model has been identified for the possible binding of ligands or inhibitors. This has been carried out through superimposing the predicted model on the top of three dimensional structures of the template. The active site residues that are key for ligand binding and their interaction with the ligand were also determined. The structural analysis of protein translocase subunit SecY (Rv0732) is believed to provide important insights for further docking analysis and to find out best probable drug like compounds with less side effects.

6.3 Results and Discussion

6.3.1 Sequence alignment and phylogenetic analysis

The primary sequence of protein translocase subunit SecY (Rv0732) was obtained from TubercuList [144]. The gene encodes 441 amino acids with molecular mass (Da) of 47579.3 and predicted theoretical Isoelectric point (pI) value of 9.80. The upregulation ranking of this gene is in the mid 40-60% and in the lower 0-20% group of genes. This is an indicator of level of upregulation of the gene in the dormant phase. A multiple sequence alignment among protein translocase subunit SecY (Rv0732) and five sequences which are orthologous group members from selected species; *Arabidopsis thaliana* , *Escherichia coli K12*, *Mycobacterium leprae strain TN, Oryza sativa and Wolbachia endosymbiont of Brugia malayi* has been carried out. Homology and evolutionary relationships are commonly inferred among the aligned sequences from multiple sequence alignment. The resulted alignment has been shown in Figure 6.2. Identical and similar amino acids are highlighted with different background colours.

Phylogenetic tree was generated from the resulted multiple sequences alignment of orthologus group members using Phylogeny.fr [147]. Phylogeny.fr is a platform designed for phylogenetic reconstruction and graphical representation through chains of programs. The inferred phylogenetic tree has been shown in Figure 6.3. In the resulted tree, there are two sub clusters where preprotein translocase subunit secY from *Arabidopsis thaliana* and from *Oryza sativa Japonica Group* are grouped together in the first cluster and the remaining four group members clustered together in the other group. preprotein translocase subunit secY from *Mycobacterium leprae strain TN* is the nearest for protein translocase subunit SecY (Rv0732) in the Phylogenetic tree. The corresponding protein translocase subunit SecY proteins from *Escherichia coli K12* and *Wolbachia endosymbiont of Brugia malayi* clustered together.

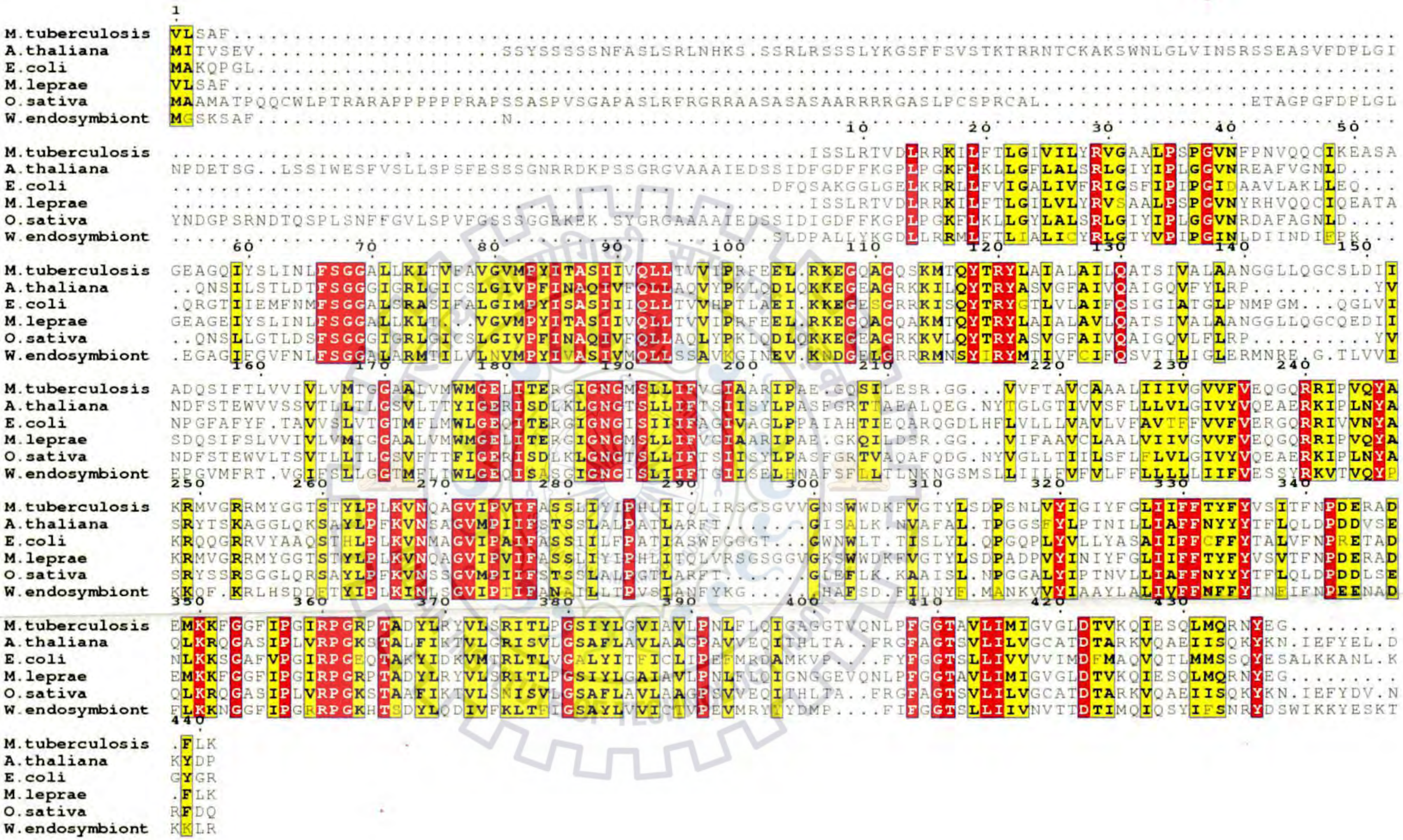


Figure 6.2 Multiple sequence alignment among orthologus group members from selected species

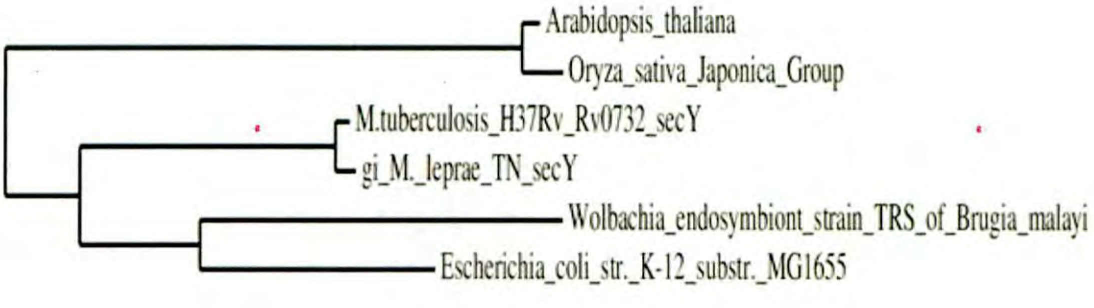

Arabidopsis_thaliana
Oryza_sativa_Japonica_Group
M.tuberculosis_H37Rv_Rv0732_secY
gi_M._leprae_TN_secY
Wolbachia_endosymbiont_strain_TRS_of_Brugia_malayi
Escherichia_coli_str._K-12_substr._MG1655

Figure 6.3 Phylogenetic tree

6.3.2 Three-dimensional Model Generation

In homology modelling, structural templates that have the highest sequence homology with the target protein are used to predict the three-dimensional structure. The sequence alignment search of protein translocase subunit SecY (Rv0732) was carried out against the three-dimensional structures deposited in Protein Data Bank (PDB) with the algorithm's default parameters (BLOSUM 62; E-threshold, 10). The search was performed using National Center for Biotechnology Information (NCBI) protein BLAST. The sequence alignment search generated hits of twenty sequences. Chain Y, Crystal Structure of Secye Translocon from Thermus thermophilus with a Fab Fragment (an accession no of 2ZJS_Y) was found to be the best alignment with 41% sequence identity and 62% sequence similarity to the target sequence. Therefore, it has been used as a template for the prediction of the structure of the target protein. A pairwise sequence alignment has been performed between the target and template sequences to visualise the conserved residues. The resulted alignment has been shown in Figure 6.4 where the conserved residues are highlighted in different background colors. In the resulted figure, the secondary structural elements of Chain Y, Crystal Structure of Secye Translocon from Thermus thermophilus with a Fab Fragment (2ZJS_Y) are shown above the aligned sequences.

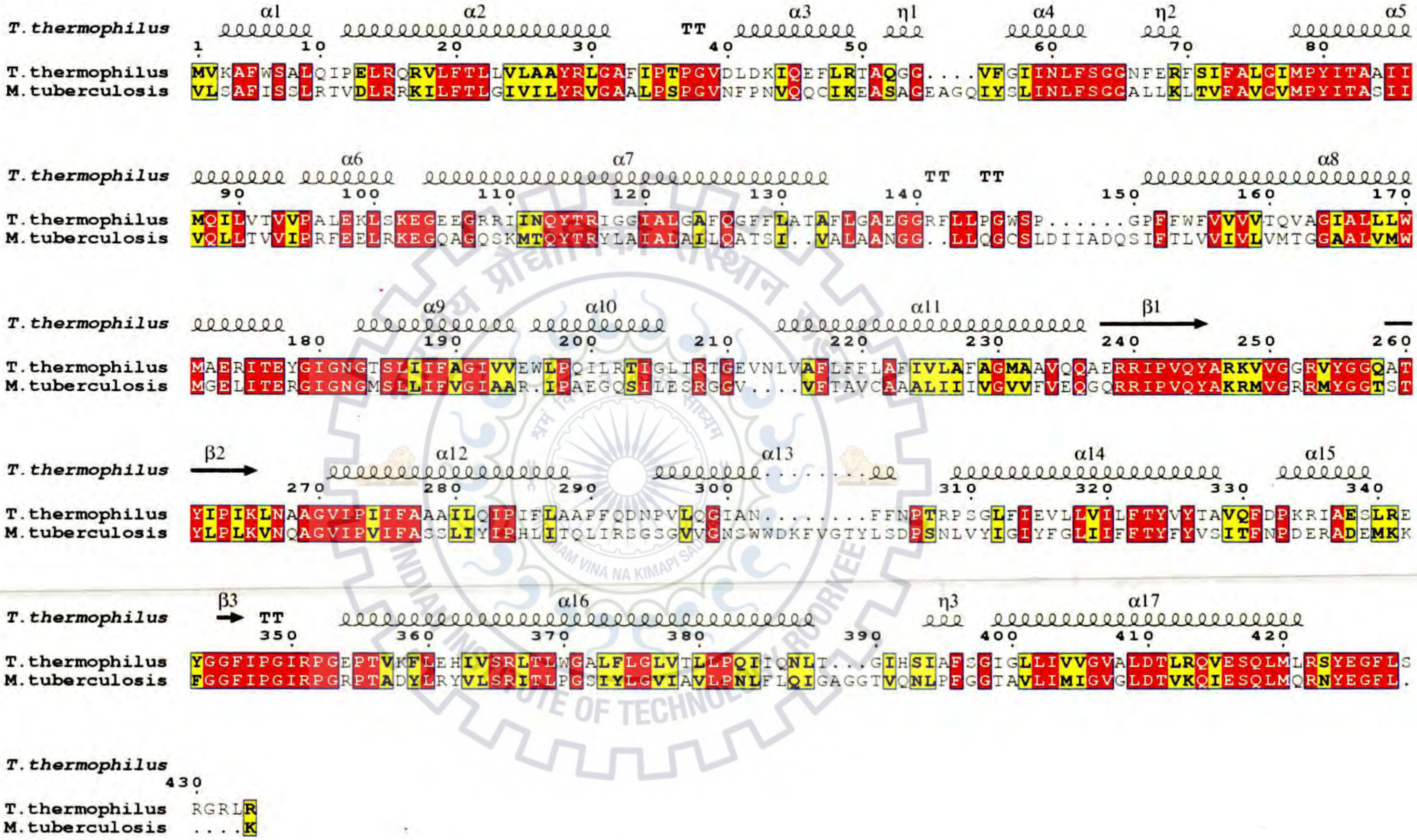

T. thermophilus
α1 α2 TT α3 η1 α4 η2 α5
1 10 20 30 40 50 60 70 80
T.thermophilus MVKAFWSALQIPELRQRVLFTLLVLAAYRLGAFIPTPGVDLDKIQEFLRTAQGG....VFGIINLFSGGNFERFSIFALGIMPYITAAII
M.tuberculosis VLSAFISSLRTVDLRRKILFTLGIVILYRVGAALPSPGVNFPNVQQCIKEASAGEAGQIYSLINLFSGGALLKLTVFAVCVMPYITASII
T. thermophilus
α6 α7 TT TT α8
90 100 110 120 130 140 150 160 170
T.thermophilus MQILVTVVPALEKLSKEGEEGRRIINQYTRIGGIALGAFQGFFLATAFLGAEGGRFLLPGWSP......GPFFWFVVVVTQVAGIALLLW
M.tuberculosis VQLLTVVIPRFEELRKEGQAGQSKMTQYTRYLAIALAILQATSI..VALAANGG..LLQGCSLDIIADQSIFTLVVIVLVMTGGAALVMW
T. thermophilus
α9 α10 α11 β1
180 190 200 210 220 230 240 250 260
T.thermophilus MAERITEYGIGNGTSLIIFAGIVVEWLPQILRTIGLIRTGEVNLVAFLFFLAFIVLAFAGMAAVQQAERRIPVQYARKVVGGRVYGGQAT
M.tuberculosis MGELITERGIGNGMSLLIFVGIAAR.IPAEGQSILESRGGV....VFTAVCAAALIIIVGVVFVEQGQRRIPVQYAKRMVGRRMYGGTST
T. thermophilus
β2 α12 α13 α14 α15
270 280 290 300 310 320 330 340
T.thermophilus YIPIKLNAAGVIPIIFAAAILQIPIFLAAPFQDNPVLQGIAN........FFNPTRPSGLFIEVLLVILFTYVYTAVQFDPKRIAESLRE
M.tuberculosis YLPLKVNQAGVIPVIFASSLIYIPHLITQLIRSGSGVVGNSWWDKFVGTYLSDPSNLVYIGIYFGLIIFFTYFYVSITFNPDERADEMKK
T. thermophilus
β3 TT α16 η3 α17
350 360 370 380 390 400 410 420
T.thermophilus YGGFIPGIRPGEPTVKFLEHIVSRLTLWGALFLGLVTLLPQIIQNLT...GIHSIAFSGIGLLIVVGVALDTLRQVESQLMLRSYEGFLS
M.tuberculosis FGGFIPGIRPGRPTADYLRYVLSRITLPGSIYLGVIAVLPNLFLQIGAGGTVQNLPFGGTAVLIMIGVGLDTVKQIESQLMQRNYEGFL.
T. thermophilus
430
T.thermophilus RGRLR
M.tuberculosis ....K

Figure 6.4 Pairwise sequence alignment between target and template sequences

The piarwise sequence alignment between the target and template has been an input to Modeller to predict models of the target. Five similar models of protein translocase subunit SecY (Rv0732) were generated based on the template structure. The assessment score values of the generated models have been shown in Table 6.1. There are different ways to select the best model from the generated models. One method is selecting the model with the lowest value of the Modeller objective function or the DOPE or SOAP assessment scores, or with the highest GA341 assessment score. GA34 is the same for all generated models and it is not also as good as DOPE or SOAP at distinguishing good models. None of the models have the lowest values in both objective function and DOPE scores. Hence, we have incorporated additional criteria to select a model with the highest number of residues in the allowed region. With this criteria target.B99990001.pdb has been selected since it has the lowest objective function and with highest number of residues in the allowed region. The cartoon diagrams of the template and the generated model has been shown in Figure 6.5 and Figure 6.6 respectively.

Table 6.1 Summary of successfully produced models

Name	molpdf	DOPE score	GA341 score
target.B99990001.pdb	2112.02319	-50774.47656	1.00000
target.B99990002.pdb	2195.69995	-50839.50000	1.00000
target.B99990003.pdb	2315.14819	-51244.17969	1.00000
target.B99990004.pdb	2352.78760	-50490.96484	1.00000
target.B99990005.pdb	2285.94092	51327.45703	1.00000

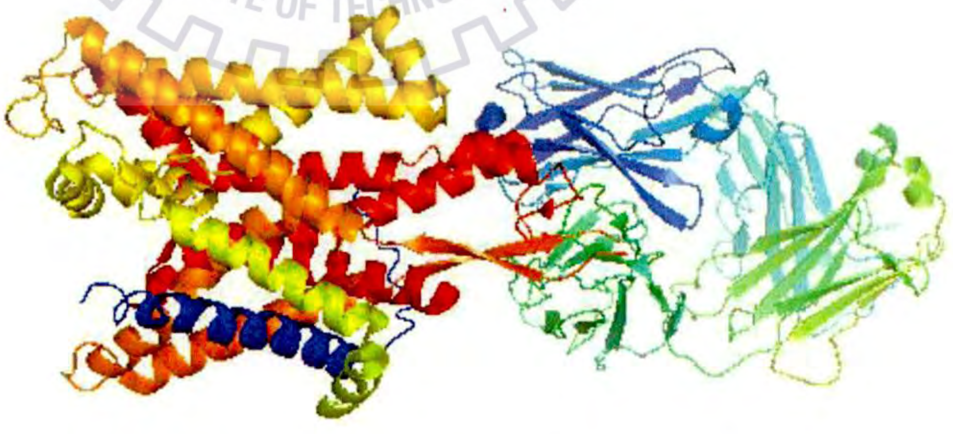


Figure 6.5 Three-dimensional structure of the template (2ZJS_Y)

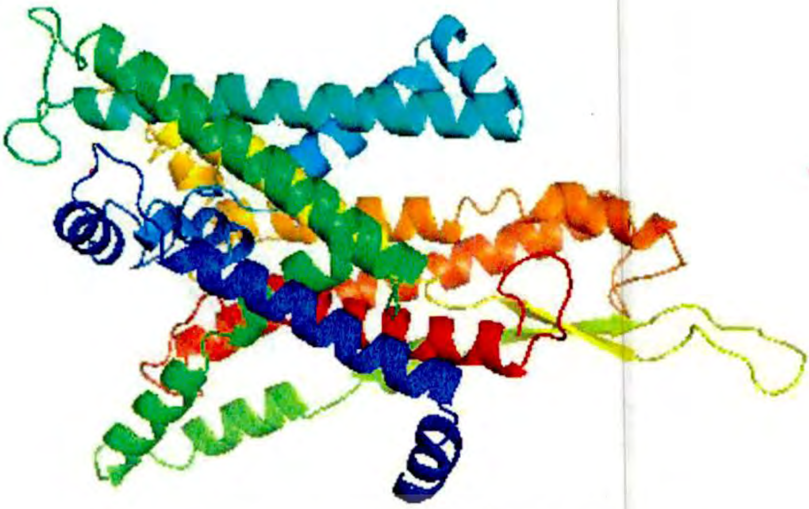
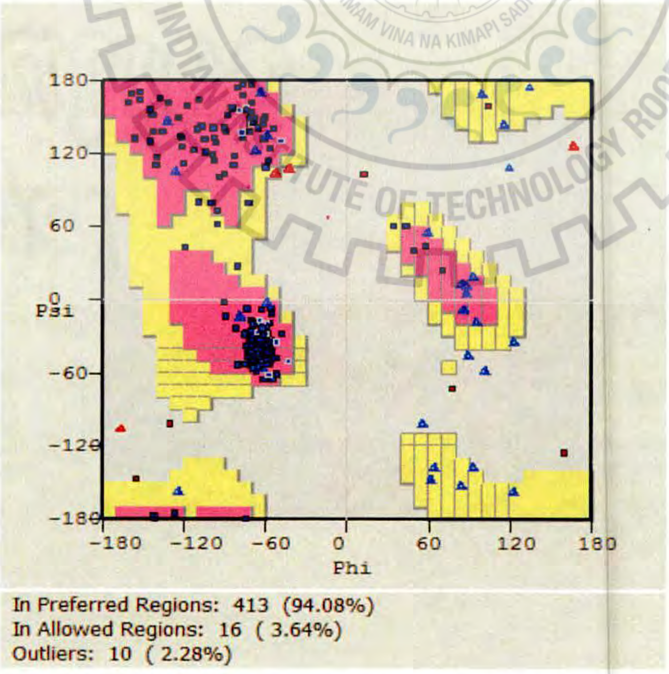


Figure 6.6 Generated model using 2ZJS_Y as template

6.3.3 Quality of the Model

The quality of the generated model was assessed using Ramachandran diagram's $\varphi - \psi$ plot. The result has been demonstrated in Figure 6.7. The majority (94.08%) of the residues lie in the most favourable region and 3.64% of the residues in the additional allowed regions. The remaining 2.28% of the residues are outliers.



In Preferred Regions: 413 (94.08%)
In Allowed Regions: 16 (3.64%)
Outliers: 10 (2.28%)

Figure 6.7 Ramachandran plot of the predicted model

6.3.4 Binding Site

In order to study the binding mode of the substrate, Zn bound structure of the generated model was identified as it has been shown in the cartoon diagram (Figure 6.8). Zn is expected to interact with a number of residues. As it has been shown in Figure 6.9, Asn 140, Arg 293 and Ser 294 bound the binding pocket. The surface view of the binding pocket is shown in Figure 6.10.

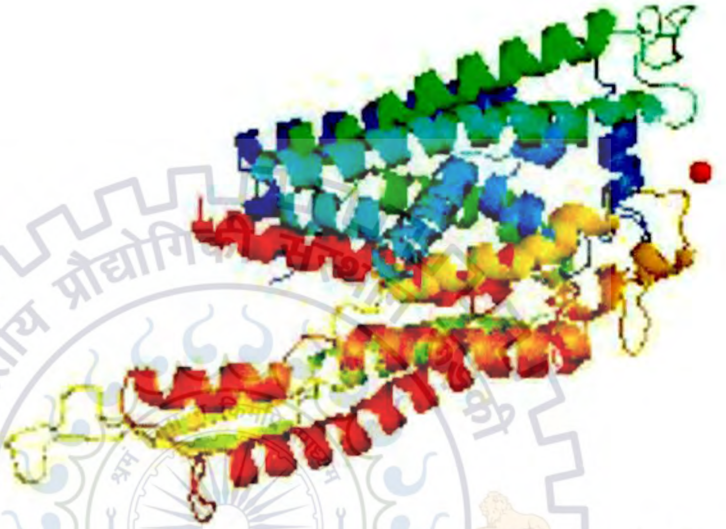


Figure 6.8 Cartoon diagram of predicted model with Zn bound to the structure

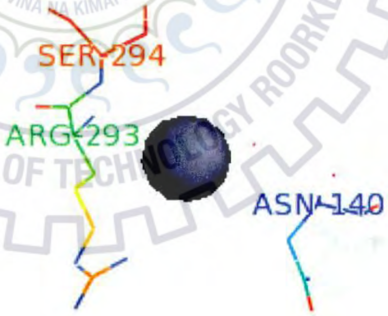


Figure 6.9 Zn binding pocket with surrounding residues

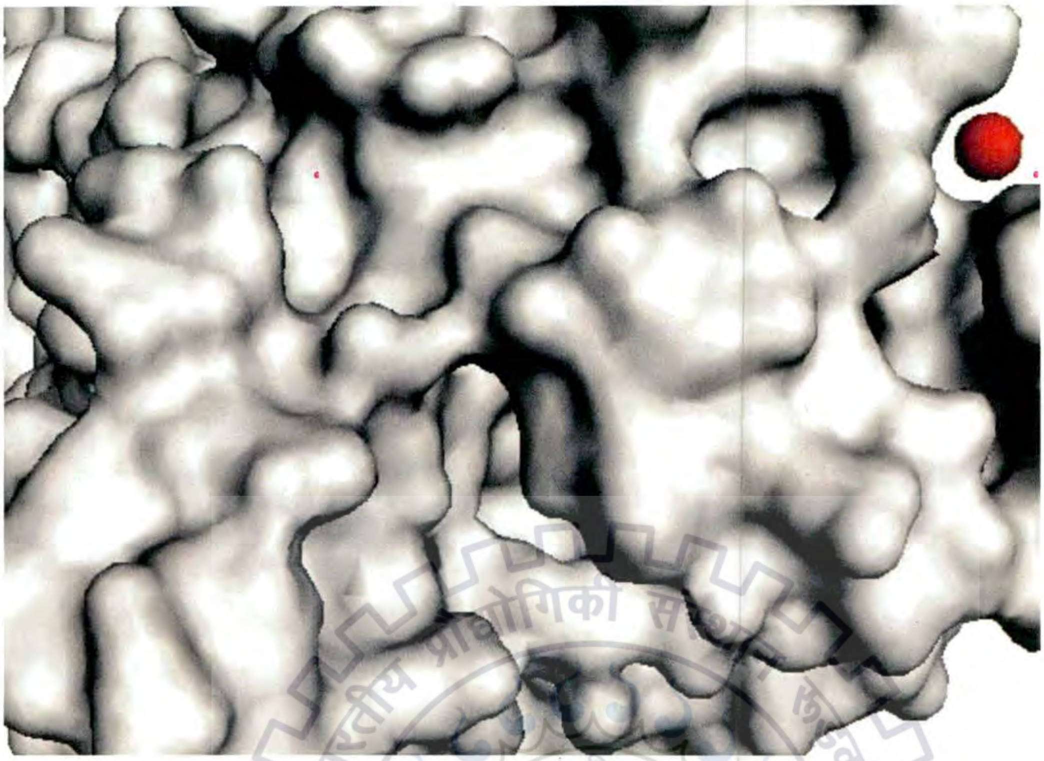


Figure 6.10 Surface view of the binding pocket

6.4 Conclusions

In this analysis, the molecular model of protein translocase subunit SecY (Rv0732) from *Mycobacterium tuberculosis H37Rv* was predicted using the crystal structure of Chain Y, Crystal Structure of Secye Translocon from Thermus Thermophilus with a Fab Fragment (an accession no of 2ZJS_Y) as template. Five models of the target were generated with Modeller 9.11. The model with lowest objective function and with the highest number of residues in the allowed region was selected. We tried to access the reliability of the model with the aid of Ramachandran plot. The diagram shows that most of the residues lie in favourable regions. Zn bound pocket of the generated model was identified to study the binding mode of the ligand and the residues which have a polar interaction with Zn were also identified. From the previous analysis, protein translocase subunit SecY (Rv0732) is highly ranked potential drug target of *Mycobacterium tuberculosis H37Rv*. There is no experimentally identified structure for the protein. Therefore, the *in silico* molecular modelling analysis carried out in this study is believed to be useful to direct the focus of synthetic chemists into designing a new drug for *Mycobacterium tuberculosis*. The model could be explored further and can be helpful for *in silico* molecular docking studies in drug design of the pathogen.

CHAPTER SEVEN

Conclusions and Future Scopes

7.1. Conclusions

Drug target identification is an important phase in the rational-based drug discovery pipeline. It is believed that the development of drugs with noble and verified drug target will significantly minimize the expensive failures of drug discovery. Identification of noble targets could be carried out either with *in vivo* and *in vitro* methods or computationally through the analysis of biomedical data and information. The effective approach is the systematic combination of both methods. The computational results could be used as input for experimental investigations to reduce the required cost and time. The experimental investigation could also be a follow up phase to validate computational results. In any of the approaches, disease mechanisms/process should be incorporated for the effectiveness of drug discovery or target identification. The system level analyses of cellular interaction networks provide detail insights about underlying principles of cellular systems. With this perspective, lists of potential targets and co-targets of *Mycobacterium tuberculosis H37Rv* have been identified through protein-protein interaction network analyses in this work.

The protein-protein interaction networks of the pathogen have been constructed using datasets retrieved from STRING database. STRING is one of the main physical and functional protein–protein interaction data sources of TB. As it has been indicated in a recent comprehensive study, *Mycobacterium tuberculosis H37Rv* protein-protein interaction datasets from STRING are of low quality by containing significant amount of false positives and false negatives. This can affect results of any analysis that is based on this dataset. To minimise this impact, only protein-protein interactions with higher scores were considered.

The potential lists of primary targets of the pathogen were identified through systematic combination of comparative genome and network centrality analysis. The comparative genome analysis was used to filter out proteins which are non-homologous with human and essential for the survival and growth to the pathogen. Non-homologous assessment has been carried out by aiming to minimize the possible host toxicity of drugs developed from the potential targets. Proteins, essential for the survival and growth of the pathogen, are primary preference drug targets for host–parasite diseases like TB. The four network centrality measures degree, closeness, betweenness and eigenvector have been used to prioritize proteins based on their respective network centrality values. A targeted attack on the proteins that are central in the constructed disease-specific network is believed to disconnect the network. A list of 137 proteins which found at the centre of gravity of interactome network was proposed as reliable list of potential primary targets for the pathogen.

The major threat for various programs of TB is the emergence and rise of drug-resistance. With the objective of systematically tackling this challenge, the primary targets were further prioritized based on their maximum flow to resistance genes. The communications to resistance genes could be disrupted by inhibiting a protein that has a higher maximum flow value. The strength of maximum flow approach is that it does not suffer from biasness of favouring shortest paths and ignoring other paths that could possibly be important in the communication of information in the cellular network since its principle is based on flow. In order to prevent the emergence of drug-resistance, we have proposed lists of proteins as potential co-targets of eight clinically used drugs in the current regime of tuberculosis treatment through maximum flow approach on drug-specific protein-protein interaction networks. The proposed potential co-targets are believed to be key players in the development of resistance by mediating information between drugs and resistance genes. These proteins can be inhibited together with the primary targets of the drugs so that the stated communication channels can be disrupted. Additionally, *in silco* structural analysis of protein translocase subunit SecY (Rv0732) has been carried out to get descriptive three-dimensional structure and identify its active site for possible protein-ligand or protein-inhibitor binding.

Validation in drug target identification is very important but difficult task due to unavailability of standard datasets. There are various sources of error that undermines the reliability of identified target such as low quality of datasets or unreliability of the computational methods. It can be done either by validating the methods used to generate the list of targets or by validating the identified targets. In these analyses, both ways have been applied where the performances of methods were assessed and the lists of potential targets have been compared with previously identified drug targets. The results showed that many of these proteins have already been reported as potential drug targets of TB. Hence, the proposed lists of potential drug targets are believed to be important in the discovery of effective medications of drug-resistance TB.

7.2. Future Scopes

Different directions have been identified for future extension of this study. We are working on upgrading the programs that have been used in computing maximum flow values in the weighted protein-protein interaction network of this thesis into a Cytoscape plug-in. The plug-in is aimed to be an application that can be used to compute maximum flows from distinguished source node to sink node of any weighted biological interaction network.

There could be a follow up detail analysis on the proposed lists of potential targets and co-targets. They can be further validated through experimental methods by constructing nock-outs. The validated targets can then be scanned against known ligand molecules for the possible identification of potential drugs specific to the targets so that they can be repurposed. This can be useful in filtering out more reliable and noble targets that can go all the way up in the drug discovery pipeline and be more successful.

The other direction where this analysis can be extended is with respect to protein-protein interaction datasets of *Mycobacterium tuberculosis H37Rv*. It has been showed that the major protein-protein interaction datasets of the pathogen are of low quality by consisting significant amount of false positives and false negatives. Hence, we were forced to use only interactions with higher scores. However, there is rapid increase in generation of interactions through computational and experimental methods. This will increase the quality and comprehensiveness of the interaction datasets. Thus, the analyses in this study can be rerun to obtain more representative and reliable targets.

The current practices on the construction of molecular interaction networks do not consider the biological contexts such as specific stimuli, tissues, cellular components and disease state which would ignore the fact that proteins favour different partners under distinct cellular conditions [148]. Integrated cellular networks could also be constructed by integrating transcription regulations and protein-protein interaction from databases which seems to be fruitful as it was effectively utilized in some investigations [46, 149]. The interactome network of the pathogen can be much more enhanced through integration of protein-protein, regulatory and signal transduction networks. These improvements on interaction networks can be incorporated in the future analyses for better results.

Moreover, multiple targets as an option can be explored more. Polypharmacolgy, One drug binds to multiple targets, is an attractive approach which is getting a lot of attention recently [112,150,151]. A reverse engineering approach like inverse docking can be applied on the proposed list of targets to predict protein targets of small molecules. The targets will be screened against the approved drug molecules with docking and the lead off-targets for further testing.

The use of multiple targets is believed to be more effective than single target drugs against mechanisms employed to fence off pathogenic attack.

Finally, the methods used in this analysis can be applied to identify potential drug targets for other high burden drug-resistance diseases like malaria.

BIBLIOGRAPHY

1. Rao VS, Srinivas K. Modern drug discovery process: an in silico approach. Journal of Bioinformatics and Sequence Analysis. 2011; 3(5):89-94.
2. Ambesi A, di Bernardo D. Computational biology and drug discovery: From single-target to network drugs. Curr Bioinform. 2006; 1: 3–13
3. Food and Drug Administration Guidance for Industry Expedited Programs for Serious Conditions. Drugs and Biologics, 2014.
4. Drews J. Drug Discovery: A historical perspective. Science. 2000, 287:1960-1964.
5. DiMasi JA, Hansen RW, Grabowski HG. The price of innovation: new estimates of drug development costs. *J Health Econ.* 2003; 22: 151-1585.
6. Sams-Dodd F. Drug discovery: selecting the optimal approach. Drug Discov Today. 2006; 11 (9-10):465-472.
7. Sams-Dodd F. Target-based drug discovery: is something wrong? Drug Discov Today. 2005; 10(2):139-147.
8. Drews J, Ryser S. Innovation deficits in the pharmaceutical industry. Drug Inf J. 1996; 30:97–108.
9. Weisbach JA, Moos WH. Diagnosing the decline of major pharmaceutical research laboratories; a prescription for drug companies. Drug Dev Res. 1995; 34:243–259.
10. Baker A, Gill J. Rethinking innovation in pharmaceutical R&D. J Commer Biotechnol. 2005; 12(1): 45–49
11. Hart CP. Finding the target after screening the phenotype. Drug Discov Today. 2005; 10(7): 513-519.
12. Yao L, Evans JA, Rzhetsky A. Novel opportunities for computational biology and sociology in drug discovery. Trends Biotechnol. 2009; 27(9):531-540.
13. Ananiadou S, Kell DB, Tsujii J. Text mining and its potential applications in systems biology. Trends Biotechnol. 2006; 24(12):571–579.
14. Rzhetsky A, Seringhaus M, Gerstein M. Seeking a new biology through text mining. Cell. 2008; 134:9–13.
15. Garten Y, Altman RB. Pharmspresso: a text mining tool for extraction of pharmacogenomic concepts and relationships from full text. BMC Bioinformatics. 2009; 10(Suppl 2):S6.
16. Dai HJ, Chang YC, Tsai RTH, Hsu WL . New challenges for biological text-mining in the next decade. J Comput Sci Technol. 2010; 25(1): 169–179.

17. Tobert JA. Lovastatin and beyond: the history of the HMG-CoA reductase inhibitors. Nat Rev Drug Discov. 2003; 2:517–526.

18. Apic G, Ignjatovic T, Boyer S, Russell RB. Illuminating drug discovery with biological pathways. FEBS Lett. 2005; 579:1872–1877.

19. Searls DB. Data integration: challenges for drug discovery. Nat Rev Drug Discov. 2005, 4(1):45-58.

20. Yan Q. Systems Biology in Drug Discovery and Development: Methods and Protocols. Humana Press, New York, NY, USA, 2010.

21. Bhardwaj A, Scaria V, Raghava GP, Lynn AM, Chandra N, Banerjee S, Raghunandanan MV, Pandey V, Taneja B, Yadav J, Dash D, Bhattacharya J, Misra A,Kumar A, Ramachandran S, Thomas Z. Open Source Drug Discovery Consortium, Brahmachari SK. Open source drug discovery--a new paradigm of collaborative research in tuberculosis drug development. Tuberculosis (Edinb). 2011; 91(5):479-86.

22. Lindsay MA. Target discovery. Nat Rev Drug Discov. 2003; 2:831–838

23. Butcher SP. Target discovery and validation in the post-genomic era. Neurochem Res. 2003; 28:367–371.

24. Smith C. Drug target validation: Hitting the target. Nature. 2003; 422:341-347.

25. Sakharkar MK, Sakharkar KR. Targetability of human disease genes. Curr Drug Discov Technol. 2007; 4:48–58

26. Yang Y, Adelstein SJ, Kassis AI. Target discovery from data mining approaches. Drug Discov Today. 2012, 14(3–4):147–154.

27. Kushwaha HR, Ghosh I. Bioinformatics Approach for Finding Target Protein in Infectious Disease. Translational Bioinformatics. 2012; 3:235–255.

28. Reddy TB1, Riley R, Wymore F, Montgomery P, DeCaprio D, Engels R, Gellesch M, Hubble J, Jen D, Jin H, Koehrsen M, Larson L, Mao M, Nitzberg M, Sisk P,Stolte C, Weiner B, White J, Zachariah ZK, Sherlock G, Galagan JE, Ball CA, Schoolnik GK. TB database: an integrated platform for tuberculosis research. Nucleic Acids Res. 2009; 37:D499-D508.

29. World Health Organization (WHO). Global tuberculosis report 2013. Geneva, Switzerland: WHO; 2014.

30. Asif SM, Asad A, Faizan A, Anjali MS, Arvind A, Neelesh K, Hirdesh K, Sanjay K. Dataset of potential targets for Mycobacterium tuberculosis H37Rv through comparative genome analysis. Bioinformation. 2009; 4:245.

31. Boshoff HI, Myers TG, Copp BR, McNeil MR, Wilson MA, Barry CE. The transcriptional responses of Mycobacterium tuberculosis to inhibitors of metabolism. J Biol Chem. 2004; 279(38):40174-84.

32. Johnson R, Streicher EM, Louw GE, Warren RM, van Helden PD, Victor TC. Drug resistance in Mycobacterium tuberculosis. Curr Issues Mol Biol. 2006; 8(2):97-111.

33. World Health Organization: Emergence of XDR-TB. WHO concern over extensive drug resistant TB strains that are virtually untreatable. Tech. rep., WHO 2006

34. Dooley SW, Jarvis WR, Martone WJ, Snider DE. Multidrug-resistant tuberculosis. Ann Intern Med. 1992; 117:257-259.

35. World Health Organization/International Union against Tuberculosis and Lung Disease Global Project on Anti-Tuberculosis Drug Resistance Surveillance. Anti-tuberculosis drug resistance in the world: report no. 3. Geneva, Switzerland: World Health Organization; 2004.

36. CDC. Emergence of Mycobacterium tuberculosis with extensive resistance to second-line drugs-worldwide, 2000–2004. MMWR Morb Mortal Wkly Rep. 2006; 55:301–305

37. Migliori GB, De Iaco G, Besozzi G, Centis R, Cirillo DM. First tuberculosis cases in Italy resistant to all tested drugs. Euro Surveill. 2007; 12(5):E070517.1.

38. Katherine Rowland (13 January 2012). "Totally drug-resistant TB emerges in India".

39. Velayati AA, Masjedi MR, Farnia P, Tabarsi P, Ghanavi J, Ziazarifi AH, Hoffner SE. Emergence of new forms of totally drug-resistant tuberculosis bacilli: super extensively drug-resistant tuberculosis or totally drug-resistant strains in iran. Chest. 2009; 136(2):420-425.

40. Udwadia ZF, Amale RA, Ajbani KK, Rodrigues C. Totally drug-resistant tuberculosis in India. Clin Infect Dis. 2012; 54(4):579-581.

41. Amir A, Rana K, Arya A, Kapoor N, Kumar H, Siddiqui MA. Mycobacterium tuberculosis H37Rv: In Silico Drug Targets Identification by Metabolic Pathways Analysis. Int J Evol Biol. 2014; 2014:284170.

42. Nunn P, Williams B, Floyd K, Dye C, Elzinga G, Raviglione M. Tuberculosis control in the era of HIV. Nat Rev Immunol. 2005; 5(10):819-826.

43. Tan YT, Tillett DJ, McKay IA. Molecular strategies for overcoming antibiotic resistance in bacteria. Mol Med Today. 2000; 6:309–314.

44. Raman K, Chandra N. Mycobacterium tuberculosis interactome analysis unravels potential pathways to drug resistance. BMC Microbiol. 2008; 8: 234.

45. Joshi RS, Jamdhade MD, Sonawane MS, Giri AP. Resistome analysis of Mycobacterium tuberculosis: Identification of aminoglycoside 2'-Nacetyltransferase (AAC) as co-target for drug desigining. Bioinformation. 2013; 9(4):174-81.

46. Yeh SH, Yeh HY, Soo VW. A network flow approach to predict drug targets from microarray data, disease genes and interactome network-case study on prostate cancer. J Clin Bioinforma. 2012;2(1):1.

47. Cheng AC, Coleman RG, Smyth KTL. Structure-based maximal affinity model predicts small-molecule druggability. Nat Biotechnol. 2007; 25:71–75.

48. Liu B, Xu J, Zou Q, Xu R, Wang X, Chen Q. Using distances between Top-n-gram and residue pairs for protein remote homology detection. BMC Bioinformatics. 2014;15 Suppl 2:S3.

49. Liu B, Wang X, Chen Q, Dong Q, Lan X. Using Amino Acid Physicochemical Distance Transformation for Fast Protein Remote Homology Detection PLoS ONE. 2012; 7(9); e46633.

50. Noble WS, Pavlidis P. Support Vector Machine and Kernel Principal Components Analysis Software Toolkit. Columbia University, New York; 2002.

51. Liu B, Wang X, Lin L, Dong Q, Wang X. A Discriminative Method for Protein Remote Homology Detection and Fold Recognition Combining Top-n-grams and Latent Semantic Analysis. BMC Bioinformatics. 2008; 9:510.

52. Liu B, Wang X, Zou Q, Dong Q, Chen Q. Protein Remote Homology Detection by Combining Chou's Pseudo Amino Acid Composition and Profile-Based Protein Representation. Mol Inf. 2013; 32:775-782.

53. Liu B, Zhang D, Xu R, Xu J, Wang X, Chen Q, Dong Q, Chou KC. Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection. Bioinformatics. 2014; 30(4):472-479.

54. Ozgür A, Vu T, Erkan G, Radev DR. Identifying gene-disease associations using centrality on a literature mined gene-interaction network. Bioinformatics. 2008; 24(13):i277-285.

55. Krauthammer M, Kaufmann CA, Gilliam TC, Rzhetsky A. Molecular triangulation: bridging linkage and molecular-network information for identifying candidate genes in Alzheimer's disease. Proc Natl Acad Sci U S A. 2004; 101(42):15148-15153.

56. Cheng D, Knox C, Young N, Stothard P, Damaraju S, Wishart DS. PolySearch: a web-based text mining system for extracting relationships between human diseases, genes,

mutations, drugs and metabolites. Nucleic Acids Res. 2008; 36 (Web Server issue), W399–W405.

57. Huang ZX, Tian HY, Hu ZF, Zhou YB, Zhao J, Yao KT. GenCLiP: a software program for clustering gene lists by literature profiling and constructing gene co-occurrence networks related to custom keywords. BMC Bioinformatics. 2008; 9: 308.

58. Prakash O, Ghosh I. Developing an antituberculosis compounds database and data mining in the search of a motif responsible for theactivity of a diverse class of antituberculosis agents. J Chem Inf Model. 2006; 46(1):17-23

59. Perry AS, Loftus B, Moroose R, Lynch TH, Hollywood D, Watson RW, Woodson K, Lawler M. In silico mining identifies IGFBP3 as a novel target of methylation in prostate cancer. Br J Cancer. 2007; 96(10):1587-1594.

60. Ryu B, Kim DS, Deluca AM, Alani RM. Comprehensive expression profiling of tumor cell lines identifies molecular signatures of melanoma progression. PLoS One. 2007; 2(7):e594.

61. Li S, Wu L, Zhang Z. Constructing biological networks through combined literature mining and microarray analysis: a LMMA approach. Bioinformatics. 2006; 22(17): 2143–2150.

62. Chou KC, Cai YD. Predicting protein–protein interactions from sequences in a hybridization space. J Proteome Res. 2006; 5(2):316– 322.

63. Hu L, Huang T, Shi X, Lu WC, Cai YD, Chou KC. Predicting Functions of Proteins in Mouse Based on Weighted Protein-Protein Interaction Network and Protein Hybrid Properties. PLoS One. 2011; 6(1): e14556.

64. Emig D, Ivliev A, Pustovalova O, Lancashire L, Bureeva S, Nikolsky Y, Bessarabova M. Drug target prediction and repositioning using an integrated network-based approach. PLoS One. 2013; 8(4):e60618.

65. Raman K, Yeturu K, Chandra N. targetTB: a target identification pipeline for Mycobacterium tuberculosis through an interactome, reactome and genome-scale structural analysis. BMC Syst Biol. 2008; 2:109.

66. Vashisht R, Mondal AK, Jain A, Shah A, Vishnoi P, Priyadarshini P, Bhattacharyya K, Rohira H, Bhat AG, Passi A, Mukherjee K, Choudhary KS, Kumar V, Arora A, Munusamy P, Subramanian A, Venkatachalam A, Gayathri S, Raj S, Chitra V, Verma K, Zaheer S, Balaganesh J, Gurusamy M, Razeeth M, Raja I,Thandapani M, Mevada V, Soni R, Rana S, Ramanna GM, Raghavan S, Subramanya SN, Kholia T, Patel

R, Bhavnani V, Chiranjeevi L, Sengupta S, Singh PK,Atray N, Gandhi S, Avasthi TS, Nisthar S, Anurag M, Sharma P, Hasija Y, Dash D, Sharma A, Scaria V, Thomas Z; OSDD Consortium, Chandra N, Brahmachari SK, Bhardwaj A. Crowd sourcing a new paradigm for interactome driven drug target identification in Mycobacterium tuberculosis. PLoS One. 2012;7(7):e39808.

67. Mazandu G, Nulder N. Generation and analysis of large-scale data-driven Mycobacterium tuberculosis functional networks for drug target identification. Adv Bioinformatics. 2011; 2011: 801478.

68. Padiadpu J, Vashisht R, Chandra N. Protein-protein interaction networks suggest different targets have different propensities for triggering drug resistance. Syst Synth Biol. 2010; 4(4):311-22.

69. Chen LC, Yeh HY, Yeh CY, Arias CR, Soo VW. Identifying co-targets to fight drug resistance based on a random walk model. BMC Syst Biol. 2012, 6:5.

70. Junker BH, Schreiber F. Analysis of Biological Networks. Volume 2. Edited by BJorn H. Junker and Falk Schreiber. John Wiley & Sons, New Jersey; 2008.

71. Pavlopoulos GA, Wegener AL, Schneider R. A survey of visualization tools for biological network analysis. BioData Min. 2008; 1:12.

72. Gursoy A, Keskin O, Nussinov R. Topological properties of protein interaction networks from a structural perspective. Biochem Soc Trans. 2008; 36(Pt 6):1398-1403.

73. Winterbach W, Mieghem PV, Reinders M, Wang H, Ridder Dd. Topology of molecular interaction networks. BMC Syst Biol. 2013; 7: 90.

74. Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, Stroedicke M, Zenkner M, Schoenherr A, Koeppen S, Timm J, Mintzlaff S, Abraham C, Bock N, Kietzmann S, Goedde A, Toksöz E, Droege A, Krobitsch S, Korn B, Birchmeier W, Lehrach H, Wanker EE. A human protein-protein interaction network: a resource for annotating the proteome. Cell. 2005; 122(6):957-968.

75. Ge H, Walhout AJ, Vidal M. Integrating ′omic′ information: A bridge between genomics and systems biology. Trends Genet. 2003; 19(10):551–560.

76. Rao VS, Srinivas K, Sujini GN, Kumar GN. Protein-protein interaction detection: methods and analysis. Int J Proteomics. 2014;2014:147648.

77. Albert R. Scale-free networks in cell biology. J Cell Sci. 2005; 118 (Pt 21):4947-4957.

78. Albert R, Jeong H, Barabasi AL. Error and attack tolerance of complex networks. Nature. 2000; 406(6794):378-382.

79. Newman M. The structure and function of complex networks. SIAM Rev. 2003; 45:167-256.

80. Ahuja RK, Magnanti TL, Orlin JB. Network Flows: Theory, Algorithms, and Applications. Prentice–Hall, Upper Saddle River, NJ; 1993.

81. Cherkassky BV, Golderg AV. On implementing push relabel method for the maximum flow problem. Algorithmica. 1994; 19:390–410.

82. Dantzig GB. Application of the Simplex Method to a Transportation Problem. In T. C. Koopmans, editor, Activity Analysis and Production and Allocation. Wiley, New York. 1951; 359–373.

83. Dantzig GB. Linear Programming and Extensions. Princeton University Press, Princeton, NJ; 1962.

84. Ford LR, Fulkerson DR. Flows in Networks. Princeton University Press, Princeton, NJ; 1962.

85. Dinitz EA. Algorithm for Solution of a Problem of Maximum Flow in Networks with Power Estimation. Soviet Math Dokl. 1970; 11:1277–1280.

86. Goldberg AV. Efficient Graph Algorithms for Sequential and Parallel Computers. PhD thesis, M.I.T., Cambridge, MA; 1987.

87. Goldberg AV, Tarjan RE. A new approach to the maximum-flow problem. J Assoc Comput Mach. 1988; 35:921–940.

88. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997; 25 (17):3389–3402.

89. Cline MS, Smoot M, Cerami E, Kuchinsky A, Landys N, Workman C, Christmas R, Avila-Campilo I, Creech M, Gross B, Hanspers K, Isserlin R, Kelley R,Killcoyne S, Lotia S, Maere S, Morris J, Ono K, Pavlovic V, Pico AR, Vailaya A, Wang PL, Adler A, Conklin BR, Hood L, Kuiper M, Sander C, Schmulevich I,Schwikowski B, Warner GJ, Ideker T, Bader GD. Integration of biological networks and gene expression data using Cytoscape. Nat Protoc. 2007; 2(10):2366-2382.

90. Tang Y, Li M, Wang J, Pan Y, Wu FX. CytoNCA: a cytoscape plugin for centrality analysis and evaluation of biological network. Biosystems. 2015; 127:67-72.

91. Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol. 1993, 234(3):779-815.

92. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN. The Protein Data Bank. Nucleic Acids Res. 2000; 28:235–242.

93. DeLano WL. PyMOL. An Open-Source Molecular Graphics Tool. DeLano Scientific, San Carlos, CA, USA; 2002.

94. Emsley P, Cowtan K. Coot: model-building tools for molecular graphics. Acta Crystallogr D Biol Crystallogr. 2004, 60(Pt 12 Pt 1):2126-2132.

95. Galagan JE, Sisk P, Stolte C, Weiner B, Koehrsen M, Wymore F, Reddy TB, Zucker JD, Engles R, Gellesch M, Hubble J, Jin H, Larson L, Mao M, Nitzberg M, White J, Zachariah ZK, Sherlock G, Ball CA, Schoolnik GK. TB database 2010: overview and update. Tuberculosis. 2010; 90(4):225-35.

96. Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, Doerks T, Julien P, Roth A, Simonovic M, Bork P, von Mering C. STRING 8–a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Res. 2009; 37: D412–D416.

97. Drug Targets Protein Database. IIT Guwahati; 2010. http://www.iitg.ernet.in/vdubey/DTPdB/index.php. Accessed 15 Jan, 2014.

98. Luo H, Lin Y, Gao F, Zhang CT. DEG 10, an update of the Database of Essential Genes that includes both protein-coding genes and non-coding genomic elements. Nucleic Acids Res. 2014; 42(Database issue):D574-D580.

99. Zhang R, Lin Y. DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes. Nucleic Acids Res. 2009; 37(Database issue):D455-D458.

100. Zhang R, Ou HY, Zhang CT. DEG, a Database of Essential Genes. Nucleic Acids Res. 2004; 32(Database issue):D271-D272.

101. Zhou H, Gao S, Nguyen NN, Fan M, Jin J, Liu B, Zhao L, Xiong G, Tan M, Li S, Wong L. Stringent homology-based prediction of H. sapiens-M. tuberculosis H37Rv protein-protein interactions. Biol Direct. 2014; 9:5.

102. Zhou H, Rezaei J, Hugo W, Gao S, Jin J, Fan M, Yong CH, Wozniak M, Wong L. Stringent DDI-based prediction of H. sapiens-M. tuberculosis H37Rv protein-protein interactions. BMC Syst Biol. 2013;7 Suppl 6:S6.

103. UniProt Consortium. The universal protein resource (UniProt). Nucleic Acids Res. 2008; 13:D190–D195.

104. Magariños MP, Carmona SJ, Crowther GJ, Ralph SA, Roos DS, Shanmugam D, Van Voorhis WC, Agüero F. TDR Targets: a chemogenomics resource for neglected diseases. Nucleic Acids Res. 2012; 40(Database issue):D1118-D1127.

105. Tomioka H, Namba K. Development of antituberculous drugs: current status and future prospects. Kekkaku. 2006; 81(12):753-774.

106. Jain A, Dixit P. Multidrug resistant to extensively drug resistant tuberculosis: What is next?; J Biosci. 2008; 33:605–616.

107. Tomioka H. Current status and perspective on drug targets in tubercle bacilli and drug design of antituberculous agents based on structure-activity relationship. Curr Pharm Des. 2014; 20(27):4305-4306.

108. Kushwaha SK, Shakya M. Protein interaction network analysis--approach for potential drug target identification in Mycobacterium tuberculosis. J Theor Biol. 2010; 262:284.

109. Von Mering C, Jensen LJ, Snel B, Hooper SD, Krupp M, Foglierini M, Jouffre N, Huynen MA, Bork P. STRING: known and predicted protein-protein associations, integrated and transferred across organisms. Nucleic Acids Res. 2005; 1(33): D433–D437.

110. Zhou H, Wong L. Comparative analysis and assessment of M. tuberculosis H37Rv protein-protein interaction datasets. BMC Genomics. 2011;12 Suppl 3:S20

111. Watts DJ, Strogatz SH. Collective dynamics of small-world networks. Nature. 1998; 393: 440–442.

112. Kinnings SL, Xie L, Fung KH, Jackson RM, Xie L, Bourne PE. The Mycobacterium tuberculosis drugome and its polypharmacological implications. PLoS Comput Biol. 2010; 6(11).

113. Sassetti CM, Boyd DH, Rubin EJ. Genes required for mycobacterial growth defined by high density mutagenesis. Mol Microbiol. 2003; 48:77–84.

114. Griffin JE, Gawronski JD, DeJesus MA, Ioerger TR, Akerley BJ, Sassetti CM. High-Resolution Phenotypic Profiling Defines Genes Essential for Mycobacterial Growth and Cholesterol Catabolism. PLoS Pathog. 2011; 7(9): e1002251.

115. Zhang YJ, Ioerger TR, Huttenhower C, Long JE, Sassetti CM, Sacchettini JC, Rubin EJ. Global Assessment of Genomic Regions Required for Growth in Mycobacterium tuberculosis. PLoS Pathog. 2012; 8(9): e1002946.

116. Melak T. and Gakkhar S. Potential non homologous protein targets of mycobacterium tuberculosis H37Rv identified from protein–protein interaction network. J. Theor Biol. 2014, 361: 152–158.

117. Mulder N, Mazandu G, Rapano H. Using Host-Pathogen Functional Interactions for Filtering Potential Drug Targets in Mycobacterium tuberculosis. J Mycobac Dis. 2013; 3: 126.

118. Forrellad MA, Klepp LI, Gioffré A, Sabio y García J, Morbidoni HR, de la Paz Santangelo M, Cataldi AA, Bigi F. Virulence factors of the Mycobacterium tuberculosis complex. Virulence. 2013; 4(1):3-66.

119. Sussman JL, Lin D, Jiang J, Manning NO, Prilusky J, Ritter O, Abola EE. Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules. Acta Crystallogr D Biol Crystallogr. 1998; 54(Pt 6 Pt 1):1078-1084.

120. Estrada E. Protein bipartivity and essentiality in the yeast protein-protein interaction network. J Proteome Res. 2006; 5(9):2177-2184.

121. Maslov S, Sneppen K. Specificity and Stability in Topology of Protein Networks. Science. 2002; 296(5569):910-913.

122. Yook S, Oltvai Z, Barabasi A. Functional and topological characterization of protein interaction networks. Proteomics. 2004; 4:928-942.

123. Newman M. A measure of betweenness centrality based on random walks. Social Networks. 2005; 27:39–54.

124. Babaei S, Hulsman M, Reinders M, de Ridder J. Detecting recurrent gene mutation in interaction network context using multi-scale graph diffusion. BMC Bioinformatics. 2013; 14:29.

125. Schroeder J, Guedes A, Duarte A. Computing the minimum cut and maximum flow of undirected graphs. Technical reports, Universidade Federal do Parana; 2004.

126. Gui WJ, Lin SQ, Chen YY, Zhang XE, Bi LJ, Jiang T. Crystal structure of DNA polymerase III β sliding clamp from Mycobacterium tuberculosis. Biochem Biophys Res Commun. 2011; 405(2):272-277.

127. Huang Q, Tonge PJ, Slayden RA, Kirikae T, Ojima I. FtsZ: a novel target for tuberculosis drug discovery. Curr Top Med Chem. 2007; 7(5):527-543.

128. Kelley B, Sharan R, Karp R, Sittler T, Root D, Ideker T. Conserved pathways within bacteria and yeast as revealed by global protein network alignment. Proc Natl Acad Sci. 2003; 100:11394–11399.

129. Nguyen L, Thompson CJ. Foundations of antibiotic resistance in bacterial physiology: the mycobacterial paradigm. Trends Microbiol. 2006; 14(7):304-312.

130. Mdluli K, Slayden RA, Zhu Y, Ramaswamy S, Pan X, Mead D, Crane DD, Musser JM, Barry CE III. Inhibition of a Mycobacterium tuberculosis β- ketoacyl ACP Synthase by Isoniazid. Science. 1998; 280:1607-1610.

131. Melak T, Gakkhar S. Maximum flow approach to prioritize potential drug targets of Mycobacterium tuberculosis H37Rv from protein-protein interaction network. Clin Transl Med. 2015; 4(1):61.

132. Gerdes SY, Scholle MD, D'Souza M, Bernal A, Baev MV, Farrell M, Kurnasov OV, Daugherty MD, Mseeh F, Polanuyer BM, Campbell JW, Anantha S, Shatalin KY, Chowdhury SA, Fonstein MY, Osterman AL. From Genetic Footprinting to Antimicrobial Drug Targets: Examples in Cofactor Biosynthetic Pathways. J Bacteriol. 2002; 184(16):4555-4572.

133. World Health Organization. "WHO Model List of Essential Medicines" . October 2013.

134. Argyrou A, Jin L, Siconilfi-Baez L, Angeletti RH, Blanchard JS. Proteome-wide profiling of isoniazid targets in Mycobacterium tuberculosis. Biochemistry. 2006; 45(47):13947-13953.

135. Bernstein J, Lott WA, Steinberg BA, Yale HL. Chemotherapy experimental tuberculosis. V. Isonicotinic acid hydrazide (nydrazid) and related compounds. Am Rev Tuberc. 1952; 65:357–364.

136. Youatt J. A review of the action of isoniazid. Am Rev Respir Dis. 1969; 99:729–749.

137. Bloom BR, Murray CJ. Tuberculosis: commentary on a reemergent killer. Science. 1992; 257:1055–1064.

138. Bourne PE, Weissig H. Structural bioinformatics, New Jersey: Wiley-Liss. Briefings in Bioinformatics. 2003; 4:509.

139. Chandonia JM, Brenner SE. The impact of structural genomics: expectations and outcomes. Science. 2006; 311(5759):347-351.

140. Kushwaha SK, Shakya M. Molecular modelling and dynamics studies of Mycobacterium tuberculosis protein RelA (Rv2583c). Int J Integ Biol. 2009; 7(3): 135-138.

141. Halgren T. New method for fast and accurate binding-site identification and analysis. Chem Biol Drug Des. 2007; 69(2):146-148.

142. Halgren TA. Identifying and characterizing binding sites and assessing druggability. J Chem Inf Model. 2009; 49(2):377-389.

143. Målen H, Berven FS, Fladmark KE, Wiker HG. Comprehensive analysis of exported proteins from Mycobacterium tuberculosis H37Rv. Proteomics. 2007; 7(10):1702-1718.

144. Lew JM, Kapopoulou A, Jones LM, Cole ST. TubercuList--10 years after. Tuberculosis (Edinb). 2011; 91(1):1-7.

145. Higgins DG, Heringa J. T-Coffee: novel method for fast and accurate multiple sequence alignment. J Mol Biol. 2000; 302(1):205-217.

146. Gouet P, Courcelle E, Stuart DI, Métoz F. ESPript: analysis of multiple sequence alignments in PostScript. Bioinformatics. 1999; 15(4):305-308.

147. Dereeper A, Guignon V, Blanc G, Audic S, Buffet S, Chevenet F, Dufayard JF, Guindon S, Lefort V, Lescot M, Claverie JM, Gascuel O. Phylogeny.fr: robust phylogenetic analysis for the non-specialist. Nucleic Acids Res. 2008; 36(Web Server issue):W465-W469.

148. Lan A, Ziv-Ukelson M, Yeger-Lotem E. A context-sensitive framework for the analysis of human signalling pathways in molecular interaction networks. Bioinformatics. 2013; 29(13):i210-6.

149. Wang CY, Chen BS. Integrated cellular network of transcription regulations and protein-protein interactions. BMC Syst Biol. 2010; 4:20.

150. Reddy AS, Zhang S. Polypharmacology: drug discovery for the future. Expert Rev Clin Pharmacol. 2013; 6(1):41-7.

151. Csermely P, Agoston V, Pongor S. The efficiency of multi-target drugs: the network approach might help drug design. Trends Pharmacol Sci. 2005; 26(4):178-82.