Investigation of Protein Interaction Networks in *Mycobacterium tuberculosis* using Computational approaches

A thesis submitted to University of Hyderabad
for the award of Doctor of Philosophy in
the Department of Biotechnology and Bioinformatics
School of Life Sciences

By
DHARMAPAL BURNE
(13LTPH01)

Department of Biotechnology and Bioinformatics
School of Life Sciences
University of Hyderabad
Hyderabad – 500046
Telangana (India)

January 2019

Investigation of Protein Interaction Networks in *Mycobacterium tuberculosis* using Computational approaches

A thesis submitted to University of Hyderabad
for the award of Doctor of Philosophy in
the Department of Biotechnology and Bioinformatics
School of Life Sciences

By
DHARMAPAL BURNE
(13LTPH01)

Department of Biotechnology and Bioinformatics
School of Life Sciences
University of Hyderabad
Hyderabad – 500046
Telangana (India)

January 2019

University of Hyderabad
School of Life Sciences
Department of Biotechnology and Bioinformatics

DECLARATION

I, **DHARMAPAL BURNE**, hereby declare that this thesis entitled **"Investigation of Protein Interaction Networks in *Mycobacterium tuberculosis* using Computational approaches"** submitted by me under the guidance and supervision of Dr. Vaibhav Vindal is a bonafide research work which is also free from plagiarism. I also declare that it has not been submitted previously in part or in full to this university or any other university or institution for the award of any degree or diploma. I hereby agree that my thesis can be deposited in Shodganga/INFLIBNET.

A report on plagiarism statistics from the university librarian is enclosed.

DHARMAPAL BURNE
(Research Scholar)
Reg. No. 13LTPH01

University of Hyderabad
School of Life Sciences
Department of Biotechnology and Bioinformatics

CERTIFICATE

This is to certify that the thesis entitled **"Investigation of Protein Interaction Networks in *Mycobacterium tuberculosis* using Computational approaches"** submitted by **Mr. DHARMAPAL BURNE** bearing registration number 13LTPH01 in partial fulfillment of requirements for the award of Doctor of Philosophy in the School of Life Sciences is a bonafide research work carried out by him under my supervision and guidance.

This thesis is free from plagiarism and has not been submitted previously in part or in full to this or any other university or institution for award of any degree or diploma.

Parts of this thesis have been

A. Publications:

Published Papers:

1. **Dhammapal Bharne**, Damuka Naresh, Vaibhav Vindal, "Inferring protein interaction network of *Mycobacterium tuberculosis H37Rv* using sequence information", Res J Life Sci Bioinform Pharm Chem Sci 2018, 4(6): 57-64.

Other Published Papers:

1. Damuka Naresh, **Dhammapal Bharne**, Paramananda Saikia, Vaibhav Vindal, "Anthraquinone rich *Cassia fistula* pod extract induces IFIT1, antiviral protein", Ind J Trad Know 2018, 17(3): 474-479.
2. Raja Polavarapu, Potshangbam Angamba Meetei, Mohit Midha, **Dhammapal Bharne**, Vaibhav Vindal, "ClosIndb: A resource for computationally derived information from clostridial genomes", Infect Genet Evol 2015, 33: 127-130.

Accepted Papers:

1. **Dhammapal Bharne**, Praveen Kant, Vaibhav Vindal, "maGUI: a graphical user interface for analysis and annotation of DNA microarray data", Curr Chem Genom Transl Med.

B. Manuscript communicated:

1. **Dhammapal Bharne**, Vaibhav Vindal, "Uncovering protein interactions network of *Mycobacterium tuberculosis* using computational approaches" *(Under review)*.
2. **Dhammapal Bharne**, Bhagyashri Tawar, Vaibhav Vindal, "Uncovering *Human-M. tuberculosis* protein interactions using computational approaches" *(Under review)*.
3. **Dhammapal Bharne**, Manasa K, Vaibhav Vindal, "DAME: A database of annotated microarray experiments" *(Under review)*.

C. Conferences:

Conference Proceedings:

1. Dharmapal Bharne, Vaibhav Vindal, "*In silico* identification of high interacting proteins in *Mycobacterium tuberculosis H37Rv*", **Journal of Proteins and Proteomics Conference Proceedings** 2014, 5(3): D-2-04.

Paper Presentations:

1. Dhammapal Bharne, Vaibhav vandal, "Inferring protein interaction network of *Mycobacterium tuberculosis H37Rv* using only sequence information" at **International Conference on**

Discrete Mathematics and its Applications to Network Science, July 7 - 10, 2018, BITS Pilani, K K Birla Goa Campus, Goa, India.

Poster Presentations:

1. Dhammapal Bharne, Vaibhav Vindal, "*In silico* analysis of protein-protein interactions in *Mycobacterium tuberculosis H37Rv*" at **Indo-German International Winter School on Pathogen Biology and Genomics**, February 23 - 27, 2015, School of Life Sciences, University of Hyderabad, India.
2. Dhammapal Bharne, Vaibhav Vindal, "maGUI: a graphical user interface for microarray data analysis" at **BioQuest**, September 23 – 24, 2015, School of Life Sciences, University of Hyderabad, India.
3. Dhammapal Bharne, Vaibhav Vindal, "Protein interaction network of *Mycobacterium tuberculosis H37Rv* using only sequence information" at **BioQuest**, October 12 – 13, 2017, School of Life Sciences, University of Hyderabad, India.
4. Dhammapal Bharne, Vaibhav Vindal, "Inferring host pathogen interactions in *Mycobacterium tuberculosis* using *in silico* two hybrid system" at **International Conference on Innovations in Pharma and Biopharma Industry**, December 20 – 22, 2017, School of Life Sciences, University of Hyderabad, India.

Further, the student has passed the following courses towards fulfillment of coursework requirements for the award of Ph.D.

Sr. No.	**Course code**	**Subject name**	**Credits**	**Pass/ Fail**
1	BT801	Seminar 1	1	Pass
2	BT802	Research Ethics & Management	2	Pass
3	BT803	Biostatistics	2	Pass
4	BT804	Analytical Techniques	3	Pass
5	BT805	Lab Work	4	Pass

This is to certify that the thesis entitled **“Investigation of Protein Interaction Networks in *Mycobacterium tuberculosis* using Computational approaches”** is a record of bonafide work done by **Mr. DHARMAPAL BURNE** (13LTPH01), a research scholar of Ph.D. program in the Department of Biotechnology and Bioinformatics, School of Life Sciences, University of Hyderabad under my guidance and supervision. The thesis has not been submitted previously in part or in full to this or any other university or institution for the award of any degree or diploma.

DR. VAIBHAV VINDAL
(Supervisor)

Head of the Department **Dean of the School**

Acknowledgement

Firstly, I would like to express my gratitude to my supervisor **Dr. Vaibhav Vindal** for his invaluable guidance and supervision throughout my research.

I thank my doctoral committee members **Dr. Nooruddin Khan** and **Dr. Prakash Prabhu** for critical advice and suggestions during my research.

I thank the present and former Heads of the Department of Biotechnology and Bioinformatics, **Prof. JSS Prakash, Prof. Padmashree, Prof. Anand Kumar Kondapi, Prof. Niyaz Ahmed** and **Prof. P Prakash Babu** and the present and former Deans of school of life sciences, **Prof. KVA Ramaiah**, **Prof. MNV Prasad, Prof. P Reddanna, Prof. M Ramanadham** and **Prof. Aparna Dutta Gupta** for providing all the research facilities in the school and making them available to us.

I would like to thank the funding agency **UGC** for the junior and senior fellowship.

I thank my fellow lab mates and others for their cooperation during my research work.

I owe my deepest gratitude to my parents and my sisters for encouraging and supporting me throughout my journey.

In the end, my stay in HCU would not have been pleasant and better without the young friends in the campus. I thank them for making this journey jovial for me.

Dedicated to Maa and Bapuji

Table of Contents

Chapter 1

Introduction

Chapter 2

Identification of Protein-Protein Interactions of *M. tuberculosis* using Computational methods

Chapter 3

Construction and Analysis of Protein Interactions Network of *M. tuberculosis*

Chapter 4

Identification of *Human* and *M. tuberculosis* Protein Interactions using Computational methods

Chapter 5

Construction and Analysis of *Human* and *M. tuberculosis* Protein Interactions Network

Summary

List of Figures

Chapter 2

Identification of Protein-Protein Interactions of *M. tuberculosis* using Computational methods

Chapter 3

Construction and Analysis of Protein Interactions Network of *M. tuberculosis*

Chapter 4

Identification of *Human* and *M. tuberculosis* Protein Interactions using Computational methods

Chapter 5

Construction and Analysis of *Human* and *M. tuberculosis* Protein Interactions Network

List of Tables

Chapter 2

Identification of Protein-Protein Interactions of *M. tuberculosis* using Computational methods

Chapter 3

Construction and Analysis of Protein Interactions Network of *M. tuberculosis*

Chapter 4

Identification of *Human* and *M. tuberculosis* Protein Interactions using Computational methods

Chapter 5

Construction and Analysis of *Human* and *M. tuberculosis* Protein Interactions Network

List of Abbreviations

TB	**Tuberculosis**
MDR	**Multi Drug Resistant**
DOB	**Directly Observation Therapy**
BGC	**Bacille Calmette-Gurein**
WHO	**World Health Organization**
PPI	**Protein-Protein Interaction**
HMI	***Human* and *M. tuberculosis* Protein Interaction**
MTB	***M. tuberculosis***
RBBH	**Reciprocal Best BLAST Hit**
BLAST	**Basic Local Alignment Search Tool**
CRAN	**Comprehensive R Archive Network**
DGE	**Differentially Expressed Genes**
GO	**Gene Ontology**
KEGG	**Kyoto Encyclopedia of Genes and Genomes**
PCA	**Principle Component Analysis**
RBF	**Radial Basis Function**
SVM	**Support Vector Machine**
I2H	***In silico* Two Hybrid system**
MSA	**Multiple Sequence Alignment**
AIDS	**Acquired Immune Deficiency Syndrome**
TLR	**Toll Like Receptor**

CHAPTER 1

Introduction

1.1 Tuberculosis

Tuberculosis (TB) is one of the prevalent infectious diseases caused by the bacterium *Mycobacterium tuberculosis*. It is one of the airborne diseases which spread from person to person through the air. TB is of two types, latent and active TB. In case of the latent TB, the pathogen infects human macrophage cells and remains in an inactive state. It is a non-contagious with no symptoms. Sometimes, it may get converted to the active TB. In case of the active TB, the pathogen multiples in the macrophage cells causing disease. It is contagious to other persons. Symptoms of the active TB include cough, fever, night sweats, weight loss, etc. TB is generally diagnosed through tuberculin skin test where tuberculin, a purified protein derivative of the pathogen, is injected in the patient just below the inside forearm. The injected site is checked after 2 to 3 days for red hard bump. If the bump swells up to specific size, the patient is diagnosed to have TB disease. In order to confirm the disease, other tests such as blood tests, chest X-rays and sputum tests are performed alongside the skin test. All the patients detected with either active or inactive TB are prescribed with medications depending upon the age, overall health and resistance to drugs. The patients detected with inactive TB may require only one kind of antibiotic while the patients with active TB may require multiple drugs. Period of treatment of the patients with drugs vary from 6 months to few years. If the patients do not take the entire TB treatment course, the pathogen may get resistance to drugs causing multi-drug resistance (MDR) TB which is difficult to be treated. Therefore patients detected with TB are provided with proper medication and correct administration. In certain cases, the patients are recommended for Directly Observation Therapy (DOT) where a healthcare worker administers the TB medication to the patients in order to ensure the full course of treatment. In the patients diagnosed with HIV, diabetes or cancer, the treatment of TB becomes harder. If the TB is left untreated, it becomes fatal affecting kidneys, brain and heart. The vast majority of the TB cases are curable with proper medications and treatments elsewise two-third of the world's population would die due to TB.

1.2 Tuberculosis History

TB is known to be present in humans from ancient days [1]. The nodules or tubercles formed in the lungs as the reason for pathology is first established by Richard Morton in 1689 AD [2]. The pathology is named as "Tuberculosis" for the first time by J. L. Schonlein in 1839 AD. On 24th March 1882 AD, Robert Koch has identified *M. tuberculosis* as the bacillus bacterium causing TB. For this work, he was awarded the Noble Prize in Physiology and Medicine in 1905 AD. Presently, the 24th March of every year is commemorated as the World TB Day. In 1906 AD, Albert Calmette and Camille Guerin have developed Bacille Calmette-Gurein (BCG) vaccine for immunization of cows against TB. In 1921, the BCG vaccine is used for immunization of humans for the first time [3-4]. Later in 1946, streptomycin, an antibiotic, is developed for making effective treatment and cure of TB [5]. Though the number of TB cases significantly reduced then, the emergence of multidrug resistant TB and resurrection of regular TB have resulted in the declaration of TB as a global healthy emergency by the World Health Organization in 1993 [6].

1.3 Tuberculosis Pathogenesis

During infection, *M. tuberculosis* reaches air sacs of alveoli in lungs [7-9]. It is identified as a foreign particle by the macrophages and forms a phagolysosome in association with lysosomes. Since the bacterium has a thick and waxy mycolic acid capsule, it is not killed by cytokines and lytic enzymes of the human immune and lysosomal cells. The macrophages are further fused with other macrophages, immune cells and fibroblasts to form granulomas. The bacterium inside the granuloma becomes dormant resulting in latent infection [10-11]. Sometimes, the bacterium starts multiplying in granuloma immediately after the infection resulting in active infection [7]. During the latent infection, the patients do not feel sick and show no symptoms [7, 1]. Further, the infection is not spread to other persons. The bacterium lives in the patients from years to decades without causing disease. If the person grows old or the tissue gets damaged, or if the person's immune system gets compromised due to infectious diseases such as HIV or due to non-infectious diseases such as cancer and diabetes, the dormant bacterium becomes active by multiplying inside the granuloma causing active TB [7, 1, 12-14]. During the active infection, the person feels sick, weak, fever and night sweats. When the bacteria present in lungs, the disease is spread to other persons through coughing, sneezing, spitting and speaking. The disease may spread to different body parts such as the kidneys, bone and brain [7, 15-16].

1.4 Tuberculosis Diagnosis and Treatment

Latent TB is first diagnosed by the Mantoux tuberculin skin test [17]. Blood samples are collected from the persons who are diagnosed positive for the skin test. The blood samples are used to perform interferon gamma release assays [18]. Positive results for both the skin test and the blood sample assay indicate that the persons have latent TB infection. Persons diagnosed with latent TB are generally treated with the drug isoniazid alone or in combination with rifampicin or rifapentine [19-20] to prevent the conversion to active TB. For the diagnosis of active TB, the persons are first examined by chest X-ray. If the positive result is diagnosed, samples of sputum, pus or a tissue biopsy are collected from the persons for the detection of acid-fast bacilli [17]. When *M. tuberculosis* is detected in the samples, the persons have active TB [21]. Therefore, the persons detected with active TB are treated with the combinations of several antibiotics to reduce the risk of the bacteria developing antibiotic resistance [1].

1.5 *Mycobacterium tuberculosis*

M. tuberculosis is the causative agent of TB disease. It is a small, aerobic and non-motile bacillus bacterium [22]. The cell wall of the bacterium has mycolic acid and high lipid content and hence it is a weakly stained gram positive bacterium [23]. It retains Ziehl-Neelsen stain; hence it is an acid fast bacillus bacterium. The bacterium multiples every 16 to 20 hours and grow only in human cells. *M. tuberculosis H37Rv* is a well-studied strain of the species that is cultured in the laboratory [24].

1.6 Tuberculosis Research

In 1890, the glycerin extract of the tubercle bacilli is announced as a remedy for TB by Robert Koch calling it as tuberculin. Though it is not effective, it is later effectively adapted as a screening test for the presence of pre-symptomatic TB [25]. In 1946, streptomycin and para-amino salicylate antibiotics are developed for the effective treatment of TB. The more effective drugs such as isoniazid and pyrazinamide are introduced in the early 1950s. Currently, BCG is the only efficient available vaccine for TB [26]. The BCG vaccine decreases the tuberculosis infection about 20 percent and the conversion of latent to active tuberculosis about 60 percent [27]. Therefore, it is the most widely used vaccine in the world to vaccinate the children against TB [1]. The immunity induced by the vaccine decreases after ten years; hence better vaccines are required to be developed.

More than half a century of anti-TB chemotherapy instigation, nearly 1.3 million deaths are reported in 2016 by WHO (World Health Organization Global TB report, 2017). The features that permit *M. tuberculosis* to persevere within the human cells allowed TB to remain one of the world's greatest killers in the 21st century. With the growing incidences and rising multi-drug resistant TB (WHO Report on Multidrug and extensively drug-resistant TB (M/XDR-TB), 2010), it is required a better understanding of *M. tuberculosis*, the causative agent of TB.

Understanding the biological processes and the mechanisms of TB is improved until recently. *In vitro* cultures of *M. tuberculosis* are not equivalent to *in vivo* cultures; hence the phenomenon of TB is observed in model organisms such as mice, guinea pigs and rabbits [28]. Mouse is the most frequently used model to study pathogenesis of *M. tuberculosis*. Recent technology has made it possible to analyze the physiology of *M. tuberculosis* directly from the infected human tissues [29-30]. Large number of molecular genetics tools is developed to make advances in understanding the *in vivo* biology of *M. tuberculosis* [31]. Several genes are detected to play important roles during *M. tuberculosis* persistence and drug tolerance. For example, pcaA gene product, a cyclopropane mycolic acid synthase enzyme, is predicted to be associated with bacterial virulence from the observation that the bacterial population increases in mice during the acute infection and defects in persistence stage [32]. inhA gene product is found to be an NADH-dependent enoyl-[ACP] reductase enzyme involved in the biosynthesis of mycolic acid of the cell [33]. Icl1 gene which encodes isocitrate lyase is depicted to be responsible for the extended survival of the bacterium in the mouse models [34]. hspR gene is detected to support the bacterial survival by encoding a transcription factor that represses the expression of Hsp70 heat shock protein [35]. rpsL gene of 30S ribosomal protein is shown to be one of the factors for bacterial virulence [36]. katG gene which encodes both catalase and peroxidase enzymes when deleted is observed to result in the bacterial resistance to Isoniazid drug [37]. pncA gene product is an pyrazinamidase or nicotinamidase enzyme mutation in which has resulted in the resistance of the bacterium to pyrazinamide drug [38]. atpE gene which encodes ATP synthase subunit C mutation, when deleted has resulted into the bacterial resistance to Bedaquiline drug. In addition, several other genes are uncovered to play crucial roles during TB infection such as rrs, rpsL, tlyA, gyrA, rpoA and rpoC, rpoB, whiB7 and rplC [39-45]. Such genes can attract as potential drug targets in order to shorten the drug treatment regime [46].

M. tuberculosis lives in human macrophages where there is a highly hostile environment including restriction of nutrients and reduction of oxygen tension [47]. Its ability to infect and persist under these conditions suggests that it has a unique mechanism of integrated responses to overcome such multiple stresses. The bacterium which infects and propagates through latency is one of the least understood aspects of TB [48]. Recently, it is revealed that the infection and persistent of *M. tuberculosis* requires the regulatory expression of multiple genes [49]. These genes form the basis for several transduction pathways and biological processes of the bacterium. Therefore, virulence and infection mechanism of *M. tuberculosis* can be well understood through the efficient identification of signal transduction systems utilized by the pathogen. Despite good information of individual genes, the behavior of the bacterium could be reliably predicted by considering a large set of genes involved in the complex biological systems [50].

With the advent of sequencing projects, the complete genome sequences of several organisms are determined. Protein coding genes and other important genome encoded features in the organisms are annotated [51]. These annotations have inclined the current research to move from the study of an individual protein molecule to large scale proteome wide studies. Proteins from these studies are shown to interact in a tangled network of metabolic, regulatory and signaling pathways of the cell. Various attempts are made to identify these interactions in *Saccharomyces cerevisiae* [52-53], *Drosophila melanogaster* [54], *Caenorhabditis elegans* [55] and *Escherichia coli* [56] for better understanding of the biological systems of the organisms. As a direct consequence of sequencing technologies, various "omics" fields are emerged such as proteomics, transcriptomics and metabolomics. Functions of the complex systems are described by the interactions among the proteins than the individual molecules [57]. The protein interactions are identified as the most ubiquitous and vital phenomena that occur in all the cellular activities including metabolic, regulatory and signaling pathways. Mechanisms that direct protein-protein interactions (PPIs) have become one of the important subjects in the recent era. Consequently, several high throughput methods are developed for the systematic detection of PPIs such as yeast two hybrid system (Y2H) [58-59], tandem affinity purification mass spectrometry [52, 60-61], phage display [62] and co-immunoprecipitation [63]. Development of microarray technologies along with sequencing projects have enabled in elucidating molecular functions and biological pathways [64-65] of a cell at the whole genome level. These developments facilitated significantly in understanding the one of the world's successful and ancient pathogen, *M. tuberculosis*.

A landmark achievement in the TB research is the determination of the whole genome sequence of *M. tuberculosis* by Cole et al., in 1998 [66]. The whole genome sequencing has revealed that *M. tuberculosis* consists of approximately 4000 genes. Comparative genomic analysis has indicated that the various genomic features influence the severity of TB in humans [67-68]. Analysis of the protein interaction data has revealed novel signaling pathways, uncovered functions of hypothetical proteins and provided new insights into the drug resistance properties of *M. tuberculosis* [69-70]. Recently, a systematic protein interaction study is accomplished through bacterial two hybrid system to provide useful means for dissevering infection mechanism and signaling pathways in *M. tuberculosis* [71]. The concerns related to rising multidrug resistant and HIV co-infection cases across the globe are effectively addressed through more predictive and interdisciplinary research [72]. Computational biology is one of the emerging fields in the interdisciplinary research where larger experimental datasets are integrated and analyzed to discover novel biological processes which have not been possible through the traditional methods [73-74]. With the ever-expanding amount of data from high-throughput experimental and sequencing techniques, a wide range of computation methods are developed for challenging genome-wide protein interactions predictions [75-77]. PPIs prediction methods using only sequence information has better prediction ability than other methods [78-79]. These methods in integration with the orthologous method are employed to infer the large scale protein interaction map of *M. tuberculosis* [80]. Pairs of proteins interpreted using computational methods are physically interacting, co-locate at the same region in the genome, co-express in the same environment, or implicate in the same signaling pathway [81]. PPIs network of such protein pairs are suitable for understanding various routes in the metabolic and signaling pathways in the pathogen. Highly connected proteins in the PPIs network are critical for inferring cellular functions [82]. Such proteins are investigated through flux balance analysis and shown to disrupt the metabolism of the pathogen significantly [83]. Therefore, the highly connected proteins can serve as potential drug targets in the pathogen [84].

In another perspective of PPIs prediction within the species, the establishment of host pathogen protein interactions in the intracellular pathogens such as *M. tuberculosis* is very essential. *M. tuberculosis* resides in macrophages, interact and influence several proteins of the human host [85-86] which allows the pathogen to persist within the host system [87-88]. Therefore, identification of various host factors would help in understanding the dynamics of host pathogen protein interactions which is now possible due to the immense opportunities offered by the genomic revolution. High throughput methods for the systematic identification of host pathogen interactions are scare. Therefore, computational methods are presented to identify cross species protein interactions that could implicate in the identification of novel and potential therapeutic targets [89]. Implementation of computation methods recently has revealed that the surface located *M. tuberculosis* proteins interact and modulate human proteins to acquire its nutrients [90]. Analysis of Human and *M. tuberculosis* protein interactions network has indicated that the pathogen targets the host immune response, phagocytic pathways as well as the proteins that interact with HIV [91]. Therefore, host-pathogen protein interactions are vital in elucidating the survival and persistence mechanism of the pathogen as well as the mechanism implicated during co-infection with other pathogens.

1.7 Current Challenges

Thanks to recent technological advances, substantial progress is achieved in TB research. However, the achievement does not have any noticeable impact on the current global trends of TB disease [72]. TB is a chronic disease which could develop over many years. During this period, the pathogen cannot be isolated. Further, a long generation time of the pathogen makes it difficult to grow *in vitro*. All these factors prolong the invention of new interventions and make the progress slow in TB research. Nevertheless, the improved interest in research and funding from various public sectors and organizations is giving optimism. Recently, the Stop TB Partnership led by the WHO has defined a global plan to eradicate TB by 2050 [92]. It is possible to accomplish only with the enlarged interdisciplinary research and development. Computational biology is one of the interdisciplinary research fields that can elucidate some of the key aspects of TB through various studies such as PPIs, host-pathogen interactions and drug discovery and development. The study of PPIs is one of the crucial approaches in understanding the virulence and infection mechanism of the pathogen. Though the attempts are made until recently to predict the PPIs network of *M. tuberculosis* [71, 80], the coverage of the predicted protein-protein interactions network is inadequate due to the implementation of only a few of the available methods. The consistency of the predicted PPIs in the interactions network is found to be lesser as different methods generated different types of interaction data. Further, the resources for PPIs of *M. tuberculosis* are limited due to limitations in high-throughput technologies [93]. Recent attempts for the prediction of protein interactions of *M. tuberculosis* with its human host has generated interactions lesser than few hundreds [90]. Biological evidences for such interactions are also lacking. Therefore, investigation of novel PPIs and the interactions with the human host for *M. tuberculosis* in the network perspectives is crucial for a systematic understanding of the pathogen.

1.8 Overview of the Thesis Work

The study of protein interactions networks will provide invaluable insights into the inner working mechanisms of cells as well as in uncovering the underlying pathways associated with the disease. Since an interactions network is a large scale biological and graphical data, it is an ideal challenge for bioinformatics research. For the biologists, the protein interactions network may unlock many secrets of the life. In the present study, various computational methods are employed to predict PPIs and the interactions network of *M. tuberculosis*. The novel PPIs are also investigated which could be valuable in understanding the survival and infection mechanism of the pathogen. Human and *M. tuberculosis* protein interactions and the interactions network are predicted using computational methods. The novel Human and *M. tuberculosis* proteins interactions and the critical proteins during infection and progression are also investigated. The present study is discussed in detail in the respective chapters (Chapters 2, 3, 4 and 5).

Chapter 2 describes the various computational methods available to predict PPIs. These methods are employed to predict protein-protein interactions in *M. tuberculosis*. Consistent PPIs are identified by considering PPIs predicted with minimum two of the computational methods. Further, a co-expression analysis is performed on the gene pairs of the consistent PPIs in order to generate confident PPIs. During this part of the study, maGUI, a graphical user interface, is developed to analyze and annotate the DNA microarray data.

Chapter 3 describes the protein interactions network of *M. tuberculosis*. The interactions network is generated from the confident PPIs of *M. tuberculosis*. High degree nodes and other topological properties are derived from the interactions network. Further, a set of novel PPIs is predicted and it is demonstrated using support vector classification approach.

Chapter 4 discusses the computational methods available for the prediction of host-pathogen protein interactions. The computational methods are employed to predict *Human (Homo sapiens)* and *M. tuberculosis* protein interactions (*HMIs*). Consistently predicted *HMIs* are used to perform functional annotations and co-expression analysis in order to generate confident *HMIs*. During this process, DAME, a database of annotated microarray experiments, is developed for retrieval of the annotated microarray data.

Chapter 5 describes the *HMIs* network. The *HMIs* network is generated from the confident *HMIs*. A set of novel *M. tuberculosis* proteins in *HMIs* network is identified. Pathway analysis of Human proteins is performed to gain insights into the tuberculosis infection and progression pathways. Further, a pair of *M. tuberculosis* proteins with the potential drug target characteristic is obtained.

1.9 References

1. Lawn SD, Zumla AI, “Tuberculosis”, Lancet 2011, 378(9785): 57–72.
2. Kontic O, Vasiljevic N, Jorga J, Lakic A, Jasovic-Gasic M, “Richard Morton (1637-1698)--the distinguished physician of the 17th century”, Srp Arh Celok Lek 2009, 137(11-12): 706-709.
3. Bonah C, “The 'experimental stable' of the BCG vaccine: safety, efficacy, proof, and standards, 1921-1933”, Stud Hist Philos Biol Biomed Sci 2005, 36(4): 696–721.
4. Comstock GW, “The International Tuberculosis Campaign: a pioneering venture in mass vaccination and research”, Clin Infect Dis 1994, 19(3): 528–540.
5. Persson S, “Smallpox, Syphilis and Salvation: medical breakthroughs that changed the world”, Exisle Publishing 2010.
6. WHO Global TB Programme, “TB: a global emergency, WHO report on the TB epidemic”, Geneva: World Health Organization 1994.
7. Kumar V, Abbas AK, Fausto N, Mitchell RN, “Robbins basic pathology”, Saunders Elsevier 2007, 516–522.
8. Houben EN, Nguyen L, Pieters J, “Interaction of pathogenic mycobacteria with the host immune system”, Curr Opin Microbiol 2006, 9(1): 76–85.
9. Queval CJ, Brosch R, Simeone R, “The macrophage: a disputed fortress in the battle against *Mycobacterium tuberculosis*”, Front Microbiol 2017, 8: 2284.

10. Grosset J, "*Mycobacterium tuberculosis* in the extracellular compartment: an underestimated adversary", Antimicrob Agents Chemother 2003, 47(3): 833–836.

11. Bozzano F, Marras F, De Maria A, "Immunology of tuberculosis", Mediterr J Hematol Infect Dis 2014, 6(1): e2014027.

12. Gibson PG, Abramson M, Wood-Baker Richard, Volmink J, Hensley M, Costabel U, "Evidence-based respiratory medicine", Blackwell Publishing Ltd 2005.

13. Restrepo BI, "Convergence of the tuberculosis and diabetes epidemics: renewal of old acquaintances", Clin Infect Dis 2007, 45(4): 436–438.

14. Crowley LV, "An introduction to human disease: pathology and pathophysiology correlations", Jones and Bartlett Learning 2009.

15. Herrmann JL, Lagrange PH, "Dendritic cells and *Mycobacterium tuberculosis*: which is the Trojan horse?", Pathol Biol (Paris) 2005, 53(1): 35-40.

16. Agarwal R, Malhotra P, Awasthi A, Kakkar N, Gupta D, "Tuberculosis dilated cardiomyopathy: an under-recognized entity?", BMC Infect Dis 2005, 5(1): 29.

17. Escalante P, "In the clinic. Tuberculosis", Ann Intern Med 2009, 150(11): ITC61–614.

18. National Institute for Health and Clinical Excellence, "Clinical guideline 117: TB", London: National Institute for Health and Clinical Excellence 2011.

19. The End TB Strategy, "Latent TB infection", Geneva: World Health Organization 2018.

20. Borisov AS, Bamrah Morris S, Njie GJ, Winston CA, Burton D, Goldberg S, Yelk Woodruff R, Allen L, LoBue P, Vernon A, "Update of recommendations for use of once-weekly isoniazid-rifapentine regimen to treat latent *Mycobacterium tuberculosis* infection", MMWR Morb Mortal Wkly Rep 2018, 67(25): 723-726.

21. Special Programme for Research and Training in Tropical and Foundation for Innovative New Diagnostics, “Diagnostics for TB: global demand and market potential”, Geneva: World Health Organization 2006.
22. Mandell GL, Bennett JE, Dolin R, “Mandell, Douglas, and Bennett's principles and practice of infectious diseases”, Churchill Livingstone Elsevier 2010.
23. Madison BM, “Application of stains in clinical microbiology”, Biotech Histochem 2001, 76(3): 119–125.
24. Parish T, Stoker NG, “*Mycobacteria*: bugs and bugbears (two steps forward and one step back)”, Mol Biotechnol 1999, 13(3): 191–200.
25. Waddington K, “To stamp out ‘so terrible a malady’: bovine TB and tuberculin testing in Britain, 1890-1939”, Med Hist 48(1): 29–48.
26. McShane H, “Tuberculosis vaccines: beyond Bacille Calmette-Guerin”, Philos Trans R Soc Lond B Biol Sci 2011, 366(1579): 2782-2789.
27. Roy A, Eisenhut M, Harris RJ, Rodrigues LC, Sridhar S, Habermann S, Snell L, Mangtani P, Adetifa I, Lalvani A, Abubakar I, “Effect of BCG vaccination against *Mycobacterium tuberculosis* infection in children: systematic review and meta-analysis”, BMJ 2014, 349: g4643.
28. Walsh GP, Tan EV, dela Cruz EC, Abalos RM, Villahermosa LG, Young LJ, Cellona RV, Nazareno JB, Horwitz MA, “The Phillipine cynomolgus monkey (*Macaca fasicularis*) provides a new nonhuman primate model of tuberculosis that resembles human disease”, Nat Med 1996, 2: 430–436.
29. Abbott A, “Live lung tissue enlisted in fight against tuberculosis”, Nature 2002, 415: 823.
30. Fenhalls G, Stevens L, Moses L, Bezuidenhout J, Betts JC, Helden Pv Pv, Lukey PT, Duncan K, “In situ detection of *Mycobacterium tuberculosis* transcripts in human lung granulomas reveals differential expression in necrotic lesions”, Infect Immun 2002, 70: 6330–6338.
31. Glickman MS, Jacobs Jr WR, “Microbial pathogenesis of *Mycobacterium tuberculosis*: dawn of a discipline”, Cell 2001, 104: 477–485.

32. Glickman MS, Cox JS, Jacobs Jr WR, “A novel mycolic acidcyclopropane synthetase is required for cording, persistence, and virulence of *Mycobacterium tuberculosis*”, Mol Cell 2000, 5: 717–727.

33. Banerjee A, Dubnau E, Quemard A, Balasubramanian V, Um KS, Wilson T, Collins D, de Lisle G, Jacobs WR Jr, “inhA, a gene encoding a target for isoniazid and ethionamide in *Mycobacterium tuberculosis*”, Science 1994, 263(5144): 227-230.

34. McKinney JD, Honer zu Bentrup K, Muñoz-Elías EJ, Miczak A, Chen B, Chan WT, Swenson D, Sacchettini JC, Jacobs WR Jr, Russell DG, “Persistence of *Mycobacterium tuberculosis* in macrophages and mice requires the glyoxylate shunt enzyme isocitrate lyase”, Nature 2000, 406: 735–738.

35. Stewart GR, Snewin VA, Walzl G, Hussell T, Tormay P, O'Gaora P, Goyal M, Betts J, Brown IN, Young DB, “Overexpression of heat-shock proteins reduces survival of *Mycobacterium tuberculosis* in the chronic phase of infection”, Nat Med 2001, 7: 732–737.

36. Gillespie SH, “Evolution of drug resistance in *Mycobacterium tuberculosis*: clinical and molecular perspective”, Antimicrob Agents Chemother 2002, 46(2): 267-274

37. Zhang Y, Heym B, Allen B, Young D, Cole S, “The catalase-peroxidase gene and isoniazid resistance of *Mycobacterium tuberculosis”,* Nature 1992, 358: 591-593.

38. Scorpio A, Zhang Y, “Mutations in pncA, a gene encoding pyrazinamidase/nicotinamidase, cause resistance to the antituberculous drug pyrazinamide in tubercle bacillus”, Nat Med 1996, 2: 662-667.

39. Alangaden GJ, Kreiswirth BN, Aouad A, Khetarpal M, Igno FR, Moghazeh SL, Manavathu EK, Lerner SA, “Mechanism of resistance to amikacin and kanamycin in *Mycobacterium tuberculosis*”, Antimicrob Agents Chemother 1998, 42: 1295-1297.

40. Maus CE, Plikaytis BB, Shinnick TM, “Molecular analysis of crossresistance to capreomycin, kanamycin, amikacin, and viomycin in *Mycobacterium tuberculosis*”, Antimicrob Agents Chemother 2005, 49: 3192-3197.

41. Perdigao J, Macedo R, Ribeiro A, Brum L, Portugal I, “Genetic characterisation of the ethambutol resistance-determining region in *Mycobacterium tuberculosis*: prevalence and significance of embB306 mutations”, Int J Antimicrob Agents 2009, 33: 334-338.

42. Comas I, Borrell S, Roetzer A, Rose G, Malla B, Kato-Maeda M, Galagan J, Niemann S, Gagneux S, “Whole-genome sequencing of rifampicin resistant *Mycobacterium tuberculosis* strains identifies compensatory mutations in RNA polymerase genes”, Nat Genet 2012, 44: 106-110.

43. Rahim Z, Nakajima C, Raqib R, Zaman K, Endtz HP, van der Zanden AG, Suzuki Y, “Molecular mechanism of rifampicin and isoniazid resistance in *Mycobacterium tuberculosis* from Bangladesh”, Tuberculosis (Edinb) 2012, 92: 529-534.

44. Reeves AZ, Campbell PJ, Sultana R, Malik S, Murray M, Plikaytis BB, Shinnick TM, Posey JE, “Aminoglycoside cross-resistance in *Mycobacterium tuberculosis* due to mutations in the 50 untranslated region of whiB7”, Antimicrob Agents Chemother 2013, 57: 1857-1865.

45. Makafe GG, Cao Y, Tan Y, Julius M, Liu Z, Wang C, Njire MM, Cai X, Liu T, Wang B, Pang W, “Oxazolidinone resistance in *Mycobacterium tuberculosis*: what is the role of cys154Arg mutation in the ribosomal protein L3?”, Antimicrob Agents Chemother 2016, 60: 3202-3206.

46. McKinney JD, “*In vivo* veritas: the search for TB drug targets goes live”, Nat Med 2000, 6: 1330–1333.

47. James PE, Grinberg OY, Michaels G, Swartz HM, “Intraphagosomal oxygen in stimulated macrophages”, J Cell Physiol 1995, 163(2): 241-247.

48. Manabe YC, Bishai WR, “Latent *Mycobacterium tuberculosis*-persistence, patience, and winning by waiting”, Nat Med 2000, 6(12):1327-1329.

49. Zahrt TC, Deretic V, "*Mycobacterium tuberculosis* signal transduction system required for persistent infections", Proc Natl Acad Sci U S A 2001, 98(22):12706-12711.
50. Anderson PW, "More is different", Science 1972, 177(4047): 393:396.
51. Pevsner J, "Bioinformatics and Functional Genomics", John Wiley & Sons Inc 2009.
52. Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, Remor M, Hofert C, Schelder M, Brajenovic M, Ruffner H, Merino A, Klein K, Hudak M, Dickson D, Rudi T, Gnau V, Bauch A, Bastuck S, Huhse B, Leutwein C, Heurtier MA, Copley RR, Edelmann A, Querfurth E, Rybin V, Drewes G, Raida M, Bouwmeester T, Bork P, Seraphin B, Kuster B, Neubauer G, Superti-Furga G, "Functional organization of the yeast proteome by systematic analysis of protein complexes", Nature 2002, 415(6868): 141-147.
53. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K, Yang L, Wolting C, Donaldson I, Schandorff S, Shewnarane J, Vo M, Taggart J, Goudreault M, Muskat B, Alfarano C, Dewar D, Lin Z, Michalickova K, Willems AR, Sassi H, Nielsen PA, Rasmussen KJ, Andersen JR, Johansen LE, Hansen LH, Jespersen H, Podtelejnikov A, Nielsen E, Crawford J, Poulsen V, Sørensen BD, Matthiesen J, Hendrickson RC, Gleeson F, Pawson T, Moran MF, Durocher D, Mann M, Hogue CW, Figeys D, Tyers M, "Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry", Nature 2002, 415(6868): 180-183.
54. Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao YL, Ooi CE, Godwin B, Vitols E, Vijayadamodar G, Pochart P, Machineni H, Welsh M, Kong Y, Zerhusen B, Malcolm R, Varrone Z, Collis A, Minto M, Burgess S, McDaniel L, Stimpson E, Spriggs F, Williams J, Neurath K, Ioime N, Agee M, Voss E, Furtak K, Renzulli R, Aanensen N, Carrolla S, Bickelhaupt E, Lazovatsky Y, DaSilva A, Zhong J, Stanyon CA, Finley RL Jr, White KP, Braverman M, Jarvie T, Gold S, Leach M, Knight J, Shimkets RA, McKenna

MP, Chant J, Rothberg JM, “A protein interaction map of *Drosophila melanogaster*”, Science 2003, 302(5651): 1727-1736.

55. Li S, Armstrong CM, Bertin N, Ge H, Milstein S, Boxem M, Vidalain PO, Han JD, Chesneau A, Hao T, Goldberg DS, Li N, Martinez M, Rual JF, Lamesch P, Xu L, Tewari M, Wong SL, Zhang LV, Berriz GF, Jacotot L, Vaglio P, Reboul J, Hirozane-Kishikawa T, Li Q, Gabel HW, Elewa A, Baumgartner B, Rose DJ, Yu H, Bosak S, Sequerra R, Fraser A, Mango SE, Saxton WM, Strome S, Van Den Heuvel S, Piano F, Vandenhaute J, Sardet C, Gerstein M, Doucette-Stamm L, Gunsalus KC, Harper JW, Cusick ME, Roth FP, Hill DE, Vidal M, “A map of the interactome network of the metazoan *C. elegans*”, Science 2004, 303(5657): 540-543.
56. Arifuzzaman M, Maeda M, Itoh A, Nishikata K, Takita C, Saito R, Ara T, Nakahigashi K, Huang HC, Hirai A, Tsuzuki K, Nakamura S, Altaf-Ul-Amin M, Oshima T, Baba T, Yamamoto N, Kawamura T, Ioka-Nakamichi T, Kitagawa M, Tomita M, Kanaya S, Wada C, Mori H, “Large-scale identification of protein-protein interaction of *Escherichia coli K-12*”, Genome Res 2006, 16(5): 686-691.
57. Pawson T, “Specificity in signal transduction: from phosphotyrosine-SH2 domain interactions to complex cellular systems”, Cell 2004, 116(2): 191-203.
58. Uetz P, Hughes RE, “Systematic and large-scale two-hybrid screens”, Curr Opin Microbiol 2000, 3(3): 303-308.
59. Fields S, Song O, “A novel genetic system to detect protein-protein interactions”, Nature 1989, 340(6230): 245-246.
60. Puig O, Caspary F, Rigaut G, Rutz B, Bouveret E, Bragado-Nilsson E, Wilm M, Séraphin B, “The tandem affinity purification (TAP) method: a general procedure of protein complex purification”, Methods 2001, 24(3): 218-229.
61. Rigaut G, Shevchenko A, Rutz B, Wilm M, Mann M, Seraphin B, “A generic protein purification method for protein complex characterization and proteome exploration”, Nat Biotechnol 1999, 17(10): 1030-1032.

62. Tong AH, Lesage G, Bader GD, Ding H, Xu H, Xin X, Young J, Berriz GF, Brost RL, Chang M, Chen Y, Cheng X, Chua G, Friesen H, Goldberg DS, Haynes J, Humphries C, He G, Hussein S, Ke L, Krogan N, Li Z, Levinson JN, Lu H, Ménard P, Munyana C, Parsons AB, Ryan O, Tonikian R, Roberts T, Sdicu AM, Shapiro J, Sheikh B, Suter B, Wong SL, Zhang LV, Zhu H, Burd CG, Munro S, Sander C, Rine J, Greenblatt J, Peter M, Bretscher A, Bell G, Roth FP, Brown GW, Andrews B, Bussey H, Boone C, "Global mapping of the yeast genetic interactions network", Science 2004, 303(5659): 808-813.

63. Sambrook J, Russell DW, "Identification of associated proteins by coimmunoprecipitation", CSH Protoc 2006, 2006(1): pdb.prot3898.

64. Bar-Joseph Z, Gerber GK, Lee TI, Rinaldi NJ, Yoo JY, Robert F, Gordon DB, Fraenkel E, Jaakkola TS, Young RA, Gifford DK, "Computational discovery of gene modules and regulatory networks", Nat Biotechnol 2003, 21(11): 1337-1342.

65. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N, "Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data", Nat Genet 2003, 34(2): 166-176.

66. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV, Eiglmeier K, Gas S, Barry CE 3rd, Tekaia F, Badcock K, Basham D, Brown D, Chillingworth T, Connor R, Davies R, Devlin K, Feltwell T, Gentles S, Hamlin N, Holroyd S, Hornsby T, Jagels K, Krogh A, McLean J, Moule S, Murphy L, Oliver K, Osborne J, Quail MA, Rajandream MA, Rogers J, Rutter S, Seeger K, Skelton J, Squares R, Squares S, Sulston JE, Taylor K, Whitehead S, Barrell BG, "Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence", Nature 1998, 393(6685): 537-544.

67. Tsolaki AG, Hirsh AE, DeRiemer K, Enciso JA, Wong MZ, Hannan M, Goguet de la Salmoniere YO, Aman K, Kato-Maeda M, Small PM, "Functional and evolutionary genomics of *Mycobacterium tuberculosis*:

insights from genomic deletions in 100 strains", Proc Natl Acad Sci U S A 2004, 101(14): 4865-4870.

68. Kato-Maeda M, Rhee JT, Gingeras TR, Salamon H, Drenkow J, Smittipat N, Small PM, "Comparing genomes within the species *Mycobacterium tuberculosis*", Genome Res 2001, 11(4): 547-554.

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

70. Cui T, Zhang L, Wang X, He ZG, "Uncovering new signaling proteins and potential drug targets through the interactome analysis of *Mycobacterium tuberculosis*", BMC Genomics 2009, 10:118.

71. Wang Y, Cui T, Zhang C, Yang M, Huang Y, Li W, Zhang L, Gao C, He Y, Li Y, Huang F, Zeng J, Huang C, Yang Q, Tian Y, Zhao C, Chen H, Zhang H, He ZG, "Global Protein-Protein interactions network in the Human Pathogen *Mycobacterium tuberculosis H37Rv*", J Proteome Res 2010, 9(12): 6665-6677.

72. Comas I, Gagneux S, "The Past and Future of tuberculosis Research", PLOS pathogen 2009, 5(10): e1000600.

73. Stuart LM, Boulais J, Charriere GM, Hennessy EJ, Brunet S, Jutras I, Goyette G, Rondeau C, Letarte S, Huang H, Ye P, Morales F, Kocks C, Bader JS, Desjardins M, Ezekowitz RAB, "A systems biology analysis of the *Drosophila* phagosome", Nature 2007, 445(7123): 95-101.

74. Querec TD, Akondy RS, Lee EK, Cao W, Nakaya HI, Teuwen D, Pirani A, Gernert K, Deng J, Marzolf B, Kennedy K, Wu H, Bennouna S, Oluoch H, Miller J, Vencio RZ, Mulligan M, Aderem A, Ahmed R, Pulendran B, "Systems biology approach predicts immunogenicity of the yellow fever vaccine in humans", Nat Immunol 2009, 10(1):116-125.

75. Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D, "Detecting protein function and protein-protein interactions from genome sequences", Science 1999, 285(5428): 751-753.

76. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO, “Assigning protein functions by comparative genome analysis: protein phylogenetic profiles”, Proc Natl Acad Sci U S A 1999, 96(8): 4285-4288.

77. Dandekar T, Snel B, Huynen M, Bork P, “Conservation of gene order: a fingerprint of proteins that physically interact”, Trends Biochem Sci 1998, 23(9): 324-328.

78. Shen J, Zhang J, Luo X, Zhu W, Yu K, Chen K, Li Y, Jiang H, “Predicting protein–protein interactions based only on sequences information”, PNAS 2007, 104(11): 4337-4341.

79. Bharne D, Naresh D, Vindal V, “Inferring protein interactions network of *Mycobacterium tuberculosis H37Rv* using sequence information”, Res J Life Sci Bioinform Pharm Chem Sci 2018, 4(6): 57-64.

80. Liu ZP, Wang J, Qiu YQ, Leung RK, Zhang XS, Tsui SK, Chen L, “Inferring a protein interaction map of *Mycobacterium tuberculosis* based on sequences and interlogs”, BMC Bioinformatics 2012, 13(7): S6.

81. 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, 33: D433-D437.

82. Barabasi AL, Oltvai ZN, “Network biology: understanding the cell's functional organization”, Nat Rev Genet 2004, 5(2): 101-113.

83. Raman K, Vashisht R, Chandra N, “Strategies for efficient disruption of metabolism in *Mycobacterium tuberculosis* from network analysis”, Mol Biosyst 2009, 5(12): 1740-1751.

84. Hase T, Tanaka H, Suzuki Y, Nakagawa S, Kitano H, “Structure of protein interactions networks and their implications on drug design”, PLoS Comput Biol 2009, 5(10): e1000550.

85. Raghavan S, Manzanillo P, Chan K, Dovey C, Cox JS, “Secreted transcription factor controls *Mycobacterium tuberculosis* virulence”, Nature 2008, 454(7205): 717-721.

86. Basu SK, Kumar D, Singh DK, Ganguly N, Siddiqui Z, Rao KV, Sharma P, "*Mycobacterium tuberculosis* secreted antigen (MTSA-10) modulates macrophage function by redox regulation of phosphatases", FEBS J 2006, 273(24): 5517-5534.

87. Kumar D, Nath L, Kamal MA, Varshney A, Jain A, Singh S, Rao KV, "Genome-wide analysis of the host intracellular network that regulates survival of *Mycobacterium tuberculosis*", Cell 2010, 140(5): 731-743.

88. Jayaswal S, Kamal MA, Dua R, Gupta S, Majumdar T, Das G, Kumar D, Rao KV, "Identification of host-dependent survival factors for intracellular *Mycobacterium tuberculosis* through an siRNA screen", PLoS Pathog 2010, 6(4): e1000839.

89. Dyer MD, Murali T, Sobral BW, "Computational prediction of host-pathogen protein–protein interactions", Bioinformatics 2007, 23: i159-i166.

90. Rapanoel HA, Mazandu GK and Mulder NJ, "Predicting and analyzing interactions between *Mycobacterium tuberculosis* and its human host", PLOS One 2013, 8: e67472.

91. Cui T, Li W, Liu L, Huang Q, He ZG, "Uncovering New Pathogen-Host Protein-Protein Interactions by Pairwise Structure Similarity", PLoS One 2016, 11(1): e0147612.

92. Stop TB Partnership, "The global plan to stop TB 2006-2015", Geneva: World Health Organization 2006.

93. Fabregat A, Sidiropoulos K, Garapati P, Gillespie M, Hausmann K, Haw R, Jassal B, Jupe S, Korninger F, McKay S, Matthews L, May B, Milacic M, Rothfels K, Shamovsky V, Webber M, Weiser J, Williams M, Wu G, Stein L, Hermjakob H, D'Eustachio P, "The Reactome pathway Knowledgebase", Nucleic Acids Res 2016, 44(D1): D481-D487.

CHAPTER 2

Identification of Protein-Protein Interactions of *M. tuberculosis* using Computational methods

2.1 Introduction

Function of a cell is governed by the flow of information among the cellular and molecular components. The flow of information takes place through PPIs, signal transductions, protein sorting and so on. Inferring of the PPIs and the interactions network could help in understanding the dynamics and regulations in the cellular systems. Several high-throughputs methods such as yeast-two hybrid system and mass spectrometry have emerged for the detection of PPIs. These methods are labor-intensive, time-consuming and expensive processes [1] and are prone to systemic errors. Therefore, computational methods are permeated to detect PPIs at the genomic level.

With the completion of genome sequencing projects for several organisms, large and diverse data sets are generated. Analysis and interpretations of such data make the computational methods an ideal approach [2]. These methods are employed to establish the genomic context of a gene in a complete genome [3-4]. Genes are denoted as a part of organized network of interacting proteins than an individual entity. Further, interacting proteins are specified by physical properties as well as co-expression, co-localization and the same metabolic pathways [5-7]. Therefore, PPIs are categorized as physical interactions such as the interactions among the proteins in a complex, or non-physical interactions such as the interactions among the proteins in the same molecular pathway, regulatory mechanism, cell membrane or the same soluble space [8]. Consequently, a series of computational methods are designed to predict PPIs such as Gene neighborhood [9], Gene fusion [10], Phylogenetic profile [11], Mirror tree [12] and Interlog [13]. These methods analyze different features at the genome level and detect pairs of proteins that have undergone a common evolutionary pressure which is reflected at various structural and functional levels. The PPIs predicted using these approaches is integrated with various biological evidences [14] to uncover the core inheritance mechanism of the complex biological and cellular process at the complete genome level.

In the present study, various computational methods are employed on the complete genome sequences of different mycobacterial species to detect PPIs of *M. tuberculosis H37Rv* (*MTB*). The PPIs, which are consistently predicted by multiple methods, are considered. The consistent PPIs are integrated with the vast amount of expression data [15] to obtain the confident PPIs. These PPIs are further investigated to understand the survival and infection mechanism of the pathogen.

2.2 Materials and Methods

2.2.1 Identification of a set of mycobacterial orthologs

Complete proteome sets of *Mycobacterium species* are retrieved from NCBI genome server (http://ncbi.nlm.nih.gov/genomes/). Following **Table 2.1** represents a set of mycobacterial genomes obtained from the NCBI genome server.

Table 2.1 Set of mycobacterial genomes

Mycobacterium abscessus bolletii 50594	*Mycobacterium intracellulare MOTT 64*	*Mycobacterium tuberculosis H37Rv*
Mycobacterium abscessus	*Mycobacterium JDM601*	*Mycobacterium tuberculosis H37Rv*
Mycobacterium africanum GM041182	*Mycobacterium JLS*	*Mycobacterium tuberculosis Haarlem3 NITR202*
Mycobacterium avium 104	*Mycobacterium kansasii ATCC 12478*	*Mycobacterium tuberculosis Haarlem*
Mycobacterium avium paratuberculosis K 10	*Mycobacterium KMS*	*Mycobacterium tuberculosis KZN 1435*
Mycobacterium avium paratuberculosis MAP4	*Mycobacterium leprae Br4923*	*Mycobacterium tuberculosis KZN 4207*
Mycobacterium bovis AF2122 97	*Mycobacterium leprae TN*	*Mycobacterium tuberculosis KZN 605*
Mycobacterium bovis BCG Korea 1168P	*Mycobacterium liflandii 128FXT*	*Mycobacterium tuberculosis RGTB327*
Mycobacterium bovis BCG Mexico	*Mycobacterium marinum M*	*Mycobacterium tuberculosis RGTB423*
Mycobacterium bovis BCG Pasteur 1173P2	*Mycobacterium massiliense GO 06*	*Mycobacterium tuberculosis*
Mycobacterium bovis BCG Tokyo 172	*Mycobacterium MCS*	*Mycobacterium tuberculosis UT205*
Mycobacterium canettii CIPT 140010059	*Mycobacterium MOTT36Y*	*Mycobacterium ulcerans Agy99*
Mycobacterium canettii CIPT 140060008	*Mycobacterium rhodesiae NBB3*	*Mycobacterium vanbaalenii PYR 1*
Mycobacterium canettii CIPT 140070008	*Mycobacterium smegmatis JS623*	*Mycobacterium VKM Ac 1815D*
Mycobacterium canettii CIPT 140070010	*Mycobacterium smegmatis MC2 155*	*Mycobacterium yongonense 05 1390*
Mycobacterium canettii CIPT 140070017	*Mycobacterium smegmatis MC2 155*	*Mycobacterium tuberculosis CTRI 2*
Mycobacterium chubuense NBB4	*Mycobacterium tuberculosis Beijing NITR203*	*Mycobacterium tuberculosis EAI5 NITR206*
Mycobacterium gilvum PYR GCK	*Mycobacterium tuberculosis CAS NITR204*	*Mycobacterium tuberculosis EAI5*
Mycobacterium gilvum Spyr1	*Mycobacterium tuberculosis CCDC5079*	*Mycobacterium tuberculosis Erdman ATCC 35801*
Mycobacterium indicus pranii MTCC 9506	*Mycobacterium tuberculosis CCDC5079*	*Mycobacterium tuberculosis F11*
Mycobacterium intracellulare ATCC 13950	*Mycobacterium tuberculosis CCDC5180*	*Mycobacterium tuberculosis H37Ra*
Mycobacterium intracellulare MOTT 02	*Mycobacterium tuberculosis CDC1551*	

Each protein of *MTB* proteome is examined for homologous protein in another genome using Basic Local Alignment Search Tool (BLAST). When a protein of *MTB* hits to a protein in another genome and vice versa with an *E* value less than 10^{-20} and the sequence coverage above 60 percent, such a hit is Reciprocal best BLAST hit (RBBH) and the protein pair is an orthologous proteins pair [16-18]. RBBH is performed on all the proteins of *MTB* with all the proteins of other genomes and a table of orthologous proteins of *MTB* is generated.

2.2.2 *In silico* methods to predict PPIs

To predict PPIs from the protein sequences, various computational methods are proposed in the literature [2]. These methods are employed in the present study to predict PPIs of *MTB* as described below.

2.2.2.1 Gene neighborhood

Gene neighborhood is one of the computational methods to identify PPIs from the genomic contexts. It is based on the concept that physically interacting genes are in close proximity in the genome [19-20]. Schematic representation of the Gene neighborhood method is shown in the following **Figure 2.1**. To employ this method, intergenic distances between two *MTB* genes and their corresponding orthologous genes are calculated. When the intergenic distances are less than or equal to 300 bps [9], gene products of the two genes are considered as an interacting proteins pair.

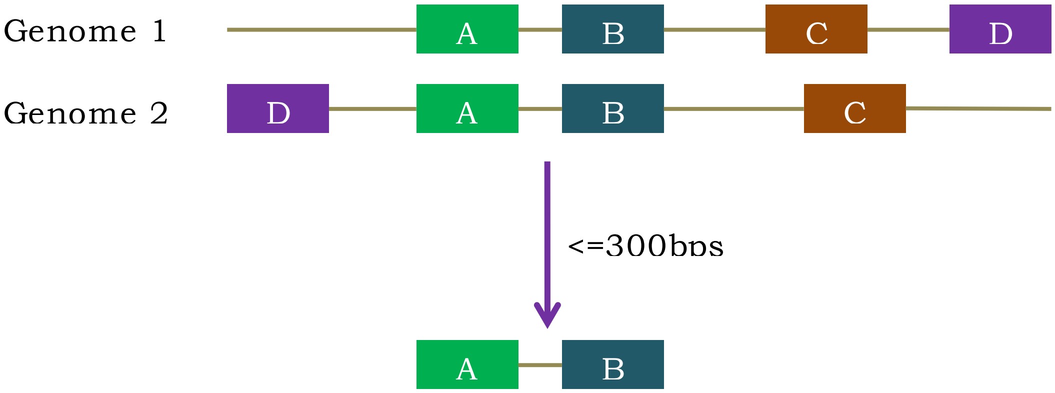
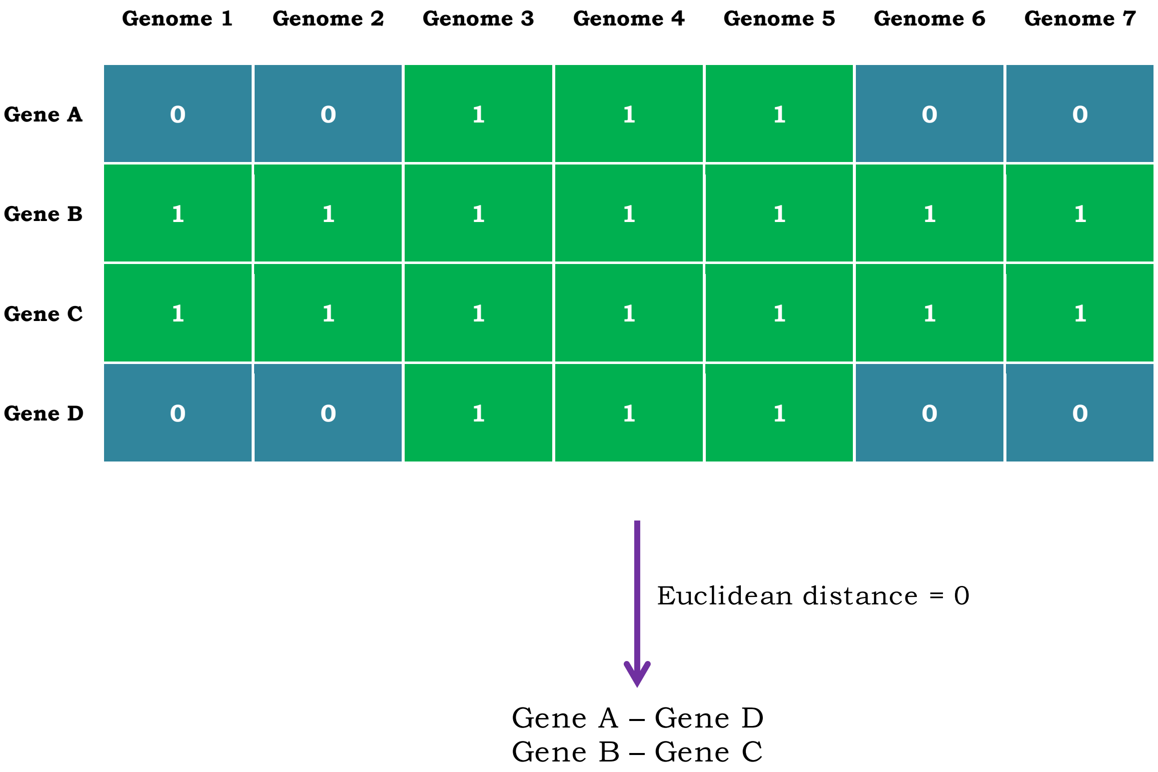


Figure 2.1 Schematic representation of the Gene neighborhood method

2.2.2.2 Phylogenetic profile

Co-occurrence of gene pairs across many genomes is the notion of the Phylogenetic profile method. Loss or gain of lineage specific genes is detected in this approach *[21-22]*. Therefore, to employ this method, orthologous table of *MTB* is used. In the table, all the orthologous proteins are substituted with a value of '1'. Thus, the profile of a *MTB* protein in the orthologous table is represented with a value of '1' for the presence of an ortholog in a genome and a value of '0' for the absence of an ortholog in the genome *[11, 23-24]*. Schematic representation of the Phylogenetic profile method is shown the following **Figure 2.2**. Profile of every protein of *MTB* is compared with the profile of every other protein of *MTB* and the Euclidean distance values are calculated. Pairs of proteins with Euclidean distance value equal to 0 are considered as interacting protein pairs *[11, 25-26]*.



	Genome 1	Genome 2	Genome 3	Genome 4	Genome 5	Genome 6	Genome 7
Gene A	0	0	1	1	1	0	0
Gene B	1	1	1	1	1	1	1
Gene C	1	1	1	1	1	1	1
Gene D	0	0	1	1	1	0	0

Figure 2.2 Schematic representation of the Phylogenetic profile method

2.2.2.3 Mirror tree

Interacting proteins tend to co-evolve [27] and the changes occurring in one protein simultaneously causes significant changes in its interacting partner. As a result, the interacting proteins have more similar phylogenetic trees than the trees of non-interacting proteins [28]. The following **Figure 2.3** represents the similarity of phylogenetic trees of two co-evolving proteins, Rv0997a and Rv2256a, which very likely to be the interacting proteins. To implement this method, the orthologs table of *MTB* is used. Pair of proteins with the number of co-incident orthologous proteins greater than or equal to 60 are compared [12, 28]. Multiple sequence alignment for each of the protein in the proteins pair is constructed and utilized to generate a distance matrix with the implementation of ClustalO software [29]. The distance matrices of two proteins are used to calculate correlation coefficient value. If the correlation coefficient value between the two proteins is minimum 0.8, the pair of proteins is considered co-evolving and interacting proteins [12, 30].

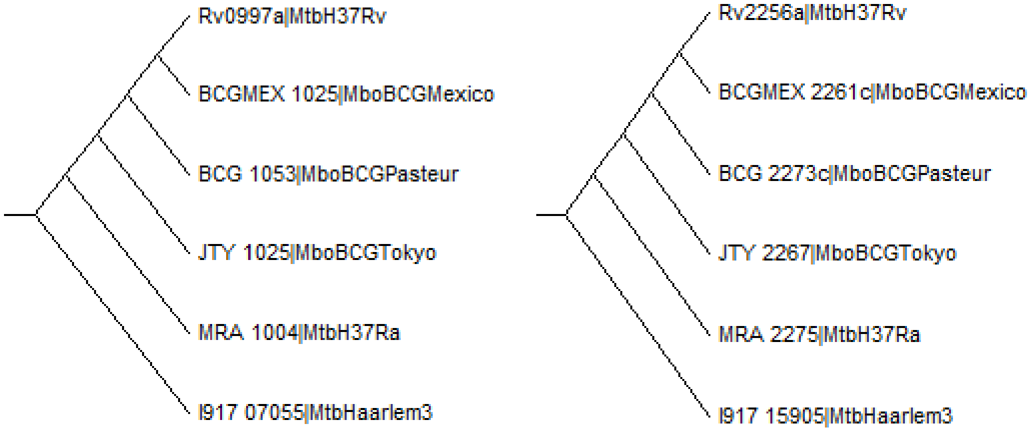


Figure 2.3 Phylogenetic trees of two co-evolving proteins, Rv0997a and Rv2256a

2.2.2.4 Gene fusion

Gene fusion method is based on the concept that two fused proteins of a genome which are independent proteins in other genomes are functionally related to each other [10]. Therefore, different genomes are compared to detect

PPIs using this approach. Schematic representation of the Gene fusion method is shown in the following **Figure 2.4**.

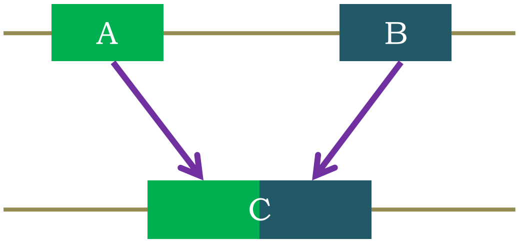


Figure 2.4 Schematic representation of the Gene fusion method

To employ this method, proteins of *MTB* are compared with proteins of a set of 40 different genomes [10, 14] shown in the following **Table 2.2**. Employing RBBH, as discussed in *Section 2.2.1*, a table of orthologous proteins and a table of paralogous proteins for *MTB* are obtained. The orthologous proteins table of *MTB* is compared with the paralogous proteins table of *MTB*. When two proteins of *MTB* are found to be not paralogs from the paralogous proteins

Table 2.2 Set of genomes employed for the Gene fusion method

Bacillus subtilis 168	*Neisseria gonorrhoeae FA 1090*
Campylobacter jejuni NCTC 11168 ATCC 700819	*Neisseria meningitidis MC58*
Clostridium botulinum A Hall	*Nocardia brasiliensis ATCC 700358*
Clostridium difficile 630	*Nocardia cyriacigeorgica GUH 2*
Clostridium perfringens 13	*Nocardia farcinica IFM 10152*
Clostridium tetani E88	*Nocardioides JS614*
Corynebacterium diphtheriae NCTC 13129	*Nocardiopsis alba ATCC BAA 2165*
Corynebacterium efficiens YS 314	*Nocardiopsis dassonvillei DSM 43111*
Corynebacterium glutamicum ATCC 13032	*Pseudomonas aeruginosa PAO1*
Corynebacterium pseudotuberculosis C231	*Rickettsia prowazekii Madrid E*
Corynebacterium resistens DSM 45100	*Rubrivivax gelatinosus IL144*
Corynebacterium terpenotabidum Y 11	*Salmonella enterica serovar Typhimurium LT2*
Corynebacterium ulcerans BR AD22	*Staphylococcus aureus NCTC 8325*
Escherichia coli K 12 substr MG1655	*Streptococcus pneumoniae R6*
Gordonia bronchialis DSM 43247	*Streptococcus pyogenes M1 GAS*
Gordonia polyisoprenivorans VH2	*Synechocystis PCC 6803*
Gordonibacter pamelaeae 7 10 1 b	*Treponema pallidum Nichols*
Helicobacter pylori 26695	*Vibrio cholerae O1 biovar El Tor N16961*
Klebsiella pneumoniae HS11286	*Xanthomonas oryzae KACC 10331*
Mycoplasma pneumoniae M129	*Yersinia pestis CO92*

table and orthologs for the same protein from the orthologous proteins table, the two proteins are considered as an interacting proteins pair and their orthologous protein as a composite of the two proteins [10].

2.2.2.5 Interlog

Many interactions in the cellular pathways and molecular mechanisms are conserved across different species and such interactions are named as "Interlogs" [31]. Schematic representation of the Interlog method is shown in the following **Figure 2.5**.

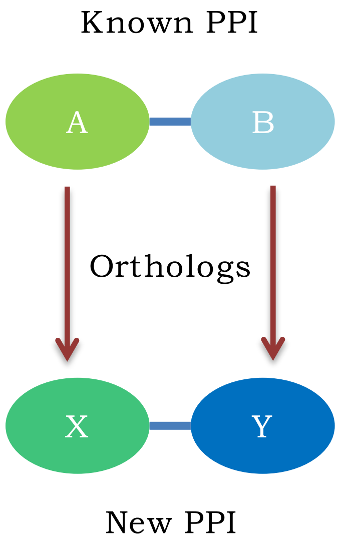


Figure 2.5 Schematic representation of the Interlog method

In order to detect *MTB* protein interlogs, PPIs are obtained from the DIP database [32] which consists of all the experimentally determined interacting proteins pairs. PPIs are also obtained from the IntAct database [33] which comprise all the molecular interactions among the proteins. Since the IntAct database stores a large volume of data, interacting proteins pairs of a set of species closely related to *MTB* are obtained. The following **Table 2.3** represents a set of genomes for PPIs obtained from the IntAct database. RBBH is performed on the protein sequences of *MTB* with the protein sequences in the PPIs retrieved from the DIP database. A table of orthologous proteins using the DIP data is obtained. It is compared with PPIs retrieved from the DIP database. Two MTB proteins that are orthologous to different proteins of a DIP interacting proteins pair, are considered as the interacting proteins pair [13, 31]. The similar procedure is employed for the data obtained from IntAct database. A table of orthologous proteins and a set of PPIs are obtained.

Table 2.3 Set of genomes for PPIs obtained from the IntAct database

Bacillus cereus	*Helicobacter pylori*	*Streptococcus agalactiae*
Campylobacter jejuni	*Klebsiella pneumonia*	*Streptococcus pneumonia*
Chlamydia pneumonia	*Mycoplasma pneumonia*	*Synechocystis sps.*
Clostridium perfringens	*Myxococcus xanthus*	*Thermosynechococcus elongates*
Desulfovibrio vulgaris	*Pseudomonas aeruginosa*	*Treponema pallidum*
Enterococcus faecalis	*Rickettsia sibirica*	*Vibrio cholera*
Escherichia coli	*Salmonella enterica*	*Yersinia pestis*
Francisella tularensis	*Staphylococcus aureus*	

2.2.3 Confident PPIs

The significance of the predicted PPIs is implied by comparing the predict PPIs with interaction pairs obtained from Wang et al. [34] and Liu et al. [35] data. Predicted PPIs are also compared with interaction pairs obtained from the STRING database [36], DIP database [32], MPIDB database [37] and the Reactome database [38] to understand the impact of the predicted PPIs. The predicted PPIs are further analyzed to generate confident PPIs by considering consistent and co-expressing genes as described below.

2.2.3.1 Consistent PPIs

To avoid false negative detections and to increase the accuracy estimations, consistent PPIs are considered [39-40]. The consistent PPIs are generated by comparing the PPIs predicted with multiple methods. Therefore, PPIs predicted by different *in silico* methods employed in the present study are compared, and sets of PPIs predicted with at least 2, 3, 4 or 5 methods are obtained.

2.2.3.2 Co-expression analysis

Consistent PPIs are used to implement co-expression analysis. In order to achieve this, microarray experimental data of *MTB* are retrieved from NCBI GEO database (http://www.ncbi.nlm.nih.gov/geo/). The experiments with the sample size greater than or equal to 10 [41] are considered for the co-expression analysis. Probes of *MTB* are not annotated; therefore, microarray data is filtered for the experiments consisting of ORF names. Such microarray data is logarithm base 2 transformed and imputed with the k-nearest neighbor method. Such data is quantile normalized and analyzed for co-expression of genes. During co-expression analysis, expression profiles of two genes are compared using Pearson correlation coefficient value. If the Pearson correlation coefficient value between two genes of a consistently predicted protein interaction pair is greater than or equal to 0.8, the two genes are considered co-expressing genes. Interactions between the two proteins which are consistently predicted with at least two *in silico* methods and whose genes are found to co-express in at least two microarray experiments are considered as confident protein interactions [41] in the present study. The following **Figure 2.6** is the flow chart for obtaining the confident PPIs.

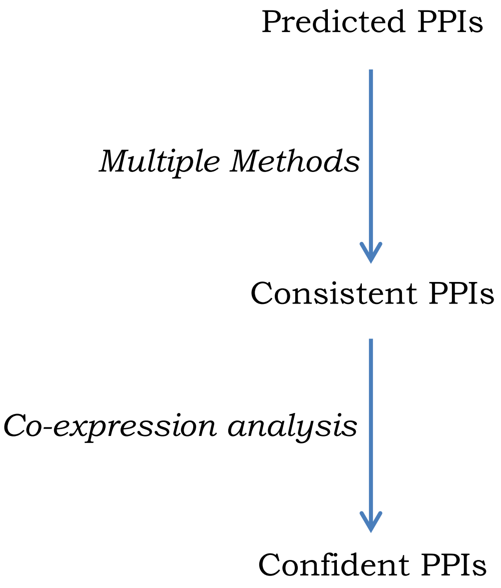


Figure 2.6 Flow-chart for the confident PPIs

2.3 Results and Discussion

2.3.1 Predicted PPIs

With the implementation of RBBH as discussed in *Section 2.2.1*, orthologous proteins of *MTB* are generated. It is observed that, of the 3906 proteins, 3,869 proteins of *MTB* are found to have orthologous protein in at least one of the other genomes. However, 37 proteins have not shown orthologous proteins in other genomes designating them as specific to *MTB*. A table of *MTB* proteins is created using the orthologous proteins. It is used to predict PPIs with the Gene neighborhood, Phylogenetic profile and the Mirror tree methods. In total, 644,572 PPIs are predicted which constitute 5,625 PPIs predicted by Gene Neighborhood method, 76,910 PPIs predicted by Phylogenetic profile and 562,037 PPIs predicted by Mirror tree method. While implementing the Gene fusion method, 104 proteins of *MTB* are observed to have paralogous proteins and 1,473 proteins are found to have orthologous proteins in other genomes. Comparing paralogous and orthologous proteins, a set of 47 PPIs is predicted. While implementing the Interlog method, a table of 26,549 orthologous proteins of *MTB* is obtained which is compared with the interaction pairs from the DIP database to generate 3,638 PPIs of *MTB*. Similarly, another table of 149,882 orthologous proteins of *MTB* is obtained which is compared with the interaction pairs from IntAct database to generate 12,386 PPIs of *MTB*. It is observed that 1900 PPIs are predicted from both the DIP and the IntAct data. Therefore, a total number of PPIs predicted using the Interlog method is 14,286 and it is represented in the following **Figure 2.7**. Overall, the number of PPIs predicted in the present study is 658,905 of which 614,490 are the unique PPIs.

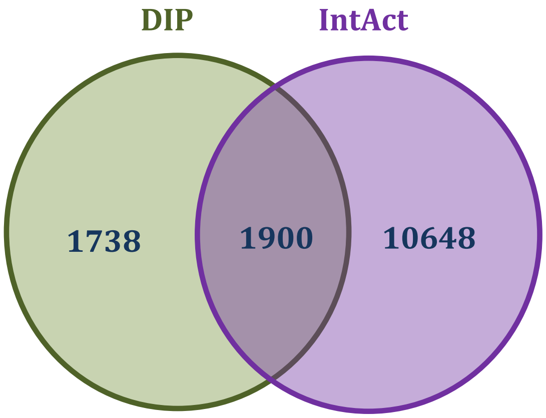


Figure 2.7 Number of PPIs predicted by the Interlog method

2.3.2 Significant PPIs

The predicted PPIs are examined in the available literature and databases as discussed in *Section 2.2.3*. It is observed that 91,967 PPIs are already known in the existing literature and databases. Thus, about 15 percent of the total predicted PPIs is already known which is in the typical range of studies related to PPIs [42-44]. The following **Table 2.4** shows the number of predicted PPIs in the known literature and databases. PPIs coverage which indicates the number of available proteins in the protein interactions network [45-46] is found to be 98.80 percent. It is higher than the coverage of the interactions network generated from Lui et al. and Wang et al. data. Therefore, the PPIs predicted in the present study have made a significant contribution towards the goals of achieving a complete protein interactions network of *MTB*.

Table 2.4 Number of overlapping PPIs in the literature and databases

Sr. No.	Source	Known PPIs in the Present Study	Total PPIs in Literature and Databases
1	Wang et al. [34]	683	8,242
2	Liu et al. [35]	5,997	43,136
3	STRING Database [36]	87,874	796,610
4	DIP Database [32]	9	19
5	MPIDB Database [37]	8	19
6	Reactome Database [38]	5	15

2.3.3 Consistent PPIs

In the total number of predicted PPIs, certain PPIs are consistently predicted with more than one *in silico* method. The following **Figure 2.8** shows the number of PPIs predicted with one or many *in silico* methods. It depicts that the number of PPIs predicted with any two methods is 43,897, any three methods is 499 and any four methods is 17 PPIs. 2 PPIs are predicted with all the five methods.

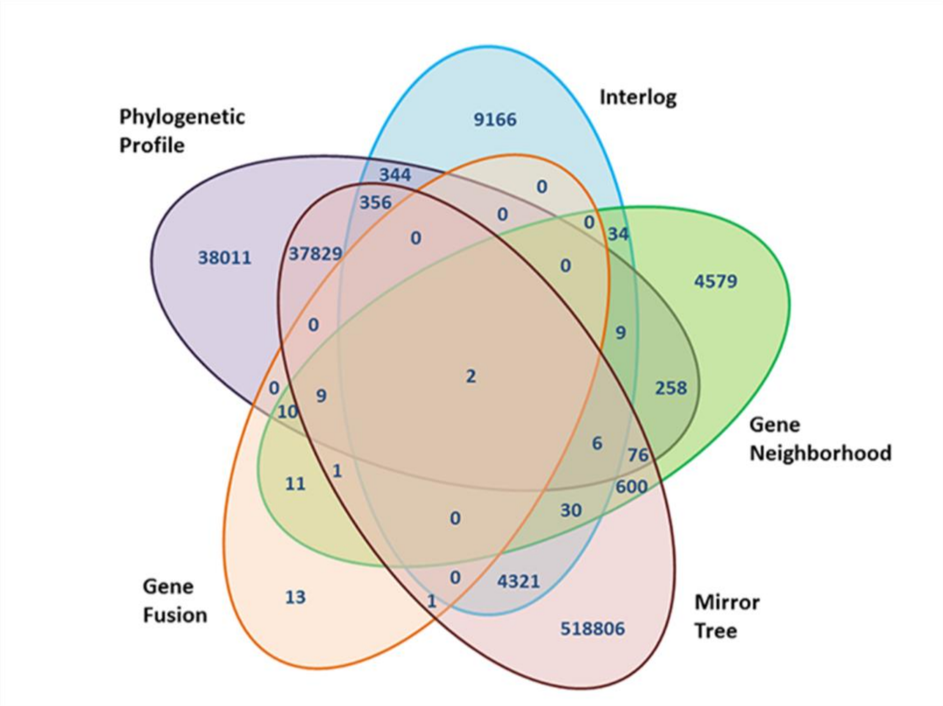


Figure 2.8 Overlap of the PPIs predicted by the computational methods

The following **Table 2.5** list the PPIs predicted with any four *in silico* methods.

Table 2.5 PPIs predicted with any four methods

Sr. No.	Protein Interaction Partner 1	Protein Interaction Partner 2
1	Rv0167	Rv0587
2	Rv0167	Rv3501c
3	Rv0168	Rv0588
4	Rv0244c	Rv1467c
5	Rv0392c	Rv1854c
6	Rv0440	Rv3417c
7	Rv0511	Rv0512
8	Rv0587	Rv3501c
9	Rv0864	Rv0865
10	Rv0951	Rv0952
11	Rv1245c	Rv3085
12	Rv1464	Rv1465
13	Rv1641	Rv1643
14	Rv1981c	Rv3048c
15	Rv2245	Rv2246
16	Rv2703	Rv2710
17	Rv2987c	Rv2988c

From **Table 2.5**, it is clear that there are few PPIs among the proteins which are of the same operon. There are coordinated interactions among the proteins belonging to different operons such as the interaction between a membrane protein Rv0167 (yrbE1A) and an integral membrane protein Rv3501c (yrbE4A). Further, PPIs of certain hypothetical proteins are revealed such as the interaction between hypothetical protein Rv0587 (yrbE2A) and the membrane protein yrbE1A, the interaction between yrbE2A and yrbE4A, and the interaction between a membrane protein Rv0168 (yrbE1B) and a hypothetical protein Rv0588 (yrbE2B). These proteins are known to play role during infection and survival of the pathogen [47-48]. A systematic view of interactions among these proteins of *MTB* is being reported for the first time from the present study. The following **Figure 2.9** shows the signal transduction system derived from the above interactions.

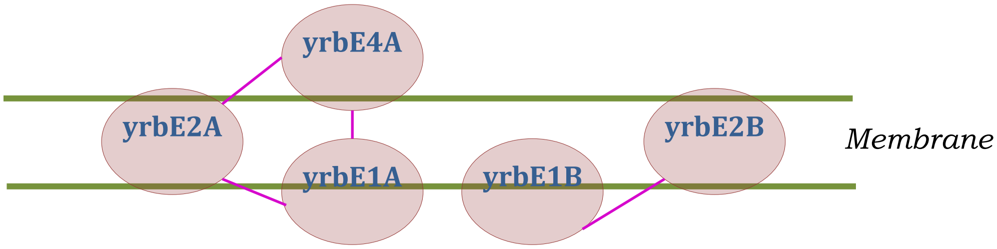


Figure 2.9 The novel signal transducing system

It is observed that the interactions of Rv0440 (groEL2) with Rv3417c (groEL1) and Rv2703 (sigA) with Rv2710 (sigB) are predicted by all the five methods employed. groEL1 and groEL2 proteins are known to play a crucial role in macrophage infections [49] and in response to damages due to heat shock [50] and reactive oxygen species [51]. The sigA and sigB RNA polymerase sigma factors of *MTB* are known to be essential in responses to complicated and varied stimuli [52-53].

2.3.4 Confident PPIs

A large number of microarray experiments related to *MTB* are available at NCBI GEO. Filtering with minimum 10 as the sample size, 75 microarray experiments are obtained. The experiments are further analyzed to extract 57 experiments that include ORF names. Pairs of proteins in the consistently predicted PPIs whose genes are co-expressing in any two of these microarray experiments are considered confident protein interaction pairs in the present study. The following **Figure 2.10** flow chart is the systematic procedure of co-expression analysis for obtaining confident *HMIs*. Using this approach, 8,243 confident PPIs are obtained. These PPIs could be essential in gaining novel insights of the TB research.

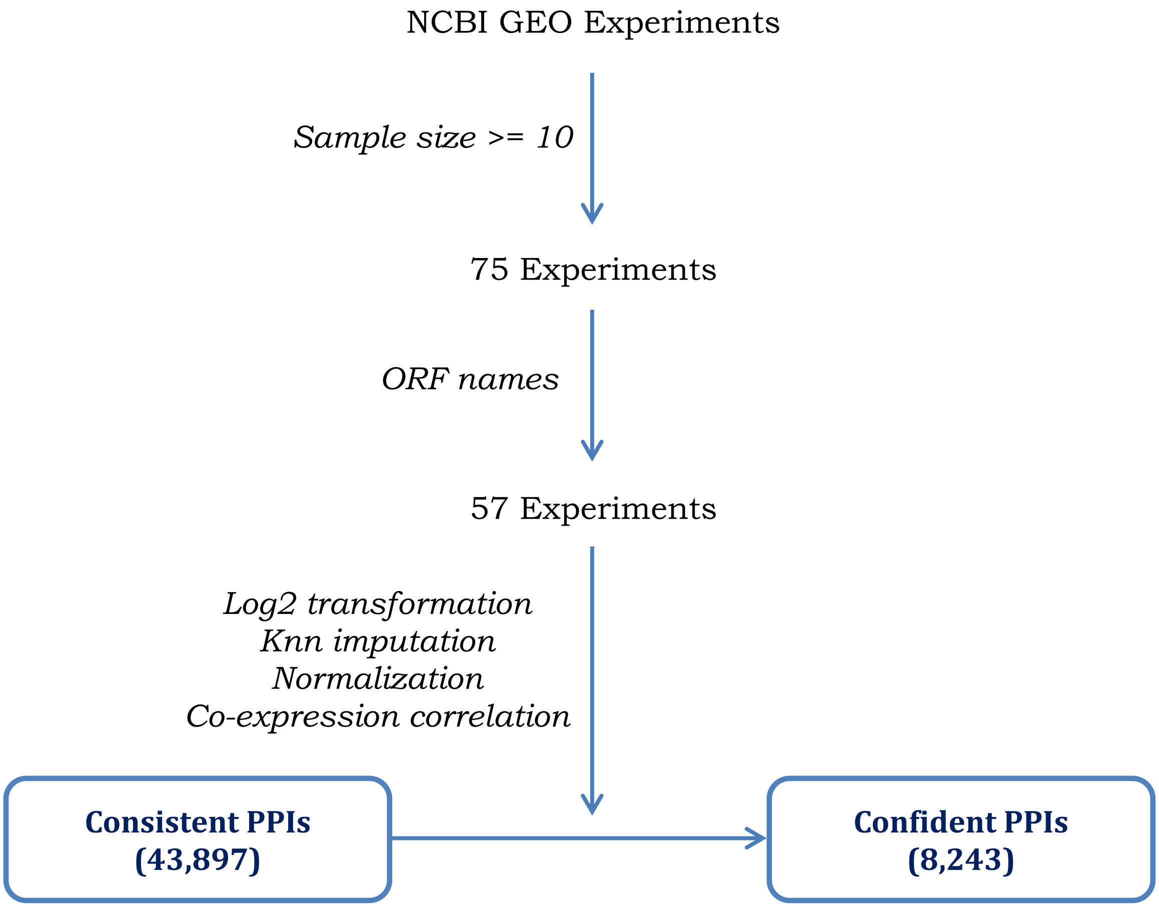


Figure 2.10 Procedure for co-expression analysis

2.3.5 The “maGUI” package

“maGUI” is a GUI (Graphical User Interface) [54] developed in-house for analysis and annotation of various types of DNA microarray data such as Affymetrix, Agilent, Illumina, Nimblegen and so on. gWidgets and dependent packages from R [55] are used to develop the GUI. Functions from the graphics and grDevices packages are employed to generate and export different types of plots during the analysis. Tcltk interface is integrated for smooth processing during export and import of graphs and tables. Several packages of Bioconductor [56] such as RSQLite, limma, beadarray, lumi, GEOquery and GEOmetadb [57-61] are used to develop the comprehensive and analytical GUI.

Visual of the GUI is organized into three regions viz., the topmost region with menus for preprocessing, analyzing and annotating the microarray data, tree region in the left hand side for the hierarchical nature of tasks that are performed on the microarray data and the leftover graphical region for viewing figures and tables produced while using the GUI. The GUI is an R package available at CRAN (Comprehensive R Archive Network) web resource. Therefore it is installed in the R environment as any other package. The microarray data is loaded from the File menu. The loaded microarray data is preprocessed, quantile normalized and quality assessed from the Pre-processing menu. The normalized data is filtered and used for the further analysis such as clustering of samples, differential gene expressions (DGE) prediction, classification of genes of interest, and so on from the Analysis menu. Annotation and visualization of the microarray data in various domains of GO [62] such as biological processes, molecular functions and cellular components, and KEGG pathways [63] is achieved from the Miscellaneous menu. The tables and figures generated during the analysis and annotation is visualized in the graphical region from the View menu and exported to local disks from the Export menu. Additional features of the maGUI includes the construction of the co-expression networks, identification of gene symbols, estimation of 2 fold change sample sizes and the prediction of protein-protein association through co-expression analysis from the two normalized data. The following **Figure 2.11** is a flow chart for the application of maGUI.

The current version of maGUI is 2.2 which can be downloaded from CRAN resource (https://cran.r-project.org/src/contrib/Archive/maGUI). It requires the R environment of version 3.0.2 or later. It is successfully tested on Linux, Windows and OSX. Help content, manual and tutorial documents can be downloaded from the site http://bif.uohyd.ac.in/maGUI/. Youtube video tutorial is available at https://www.youtube.com/watch?v=vzaoMINOGKE.

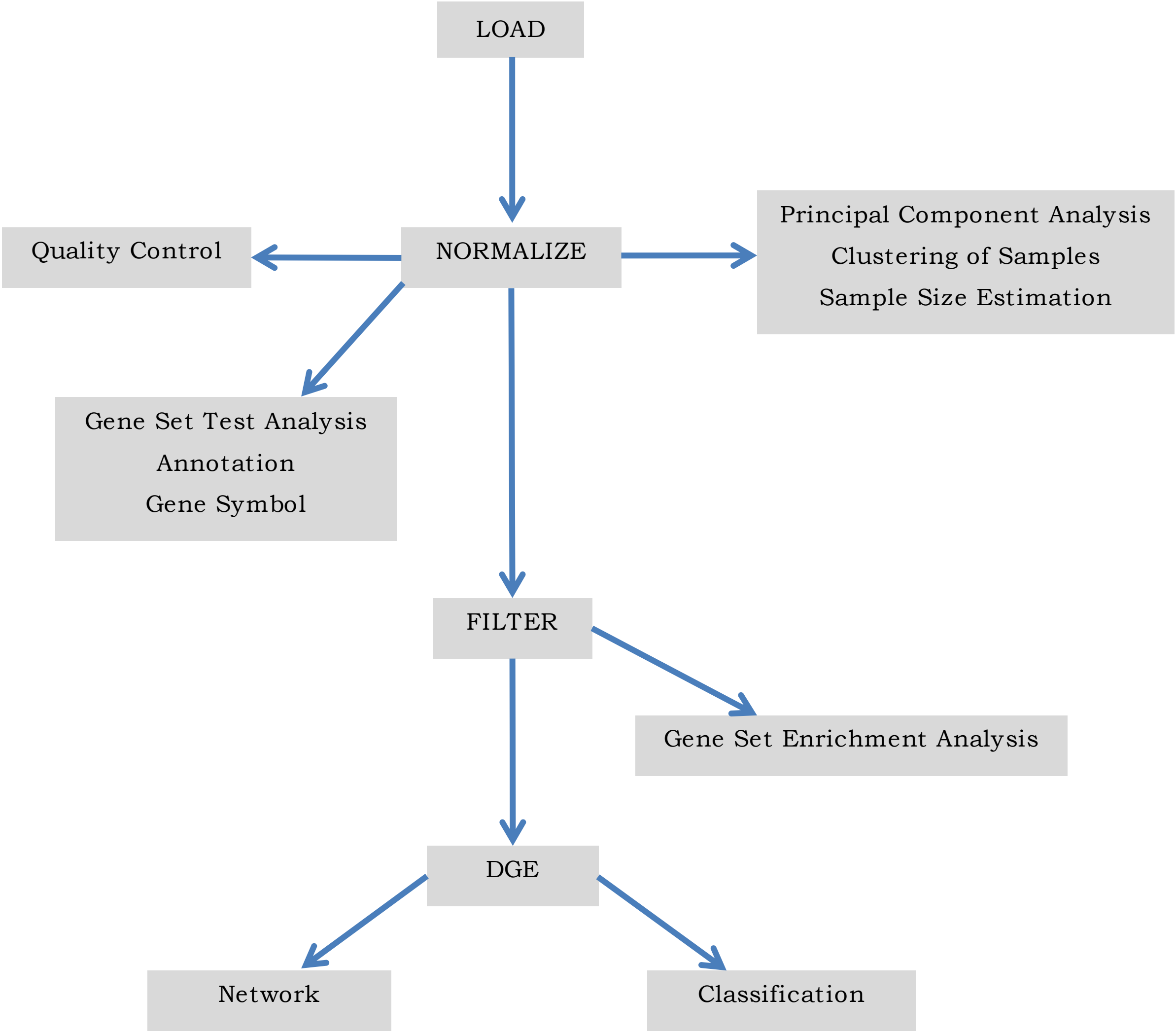


Figure 2.11 The application flow for "maGUI"

2.3.5.1 An example for maGUI application

Raw files of the microarray experiment number GSE68613 which is related to *Mus musculus* organism is downloaded from NCBI GEO. The files are extracted to a folder. The files are imported from the folder to maGUI through File → Load. The loaded microarray is normalized from Preprocessing → Normalization and the data quality is assessed from Preprocessing → Quality_Control. Principle component analysis (PCA) of normalized data is achieved through Analysis → Principal_Component_Analysis_Unsupervized. The chart of PCA obtained is in **Figure 2.12 (a)**. A cluster of samples in the experiment is visualized through Analysis → Clustering_and_Visualization_Unsupervized. Dendrogram of the clusters is shown in **Figure 2.12 (b).**

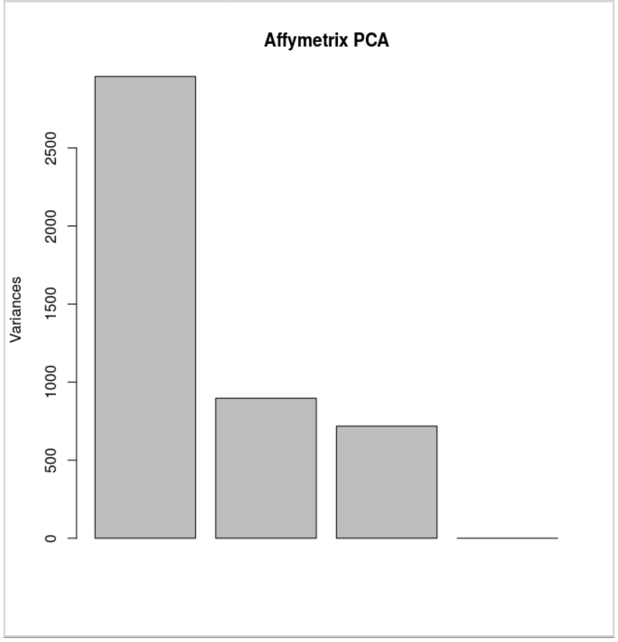


Figure 2.12 (a) PCA

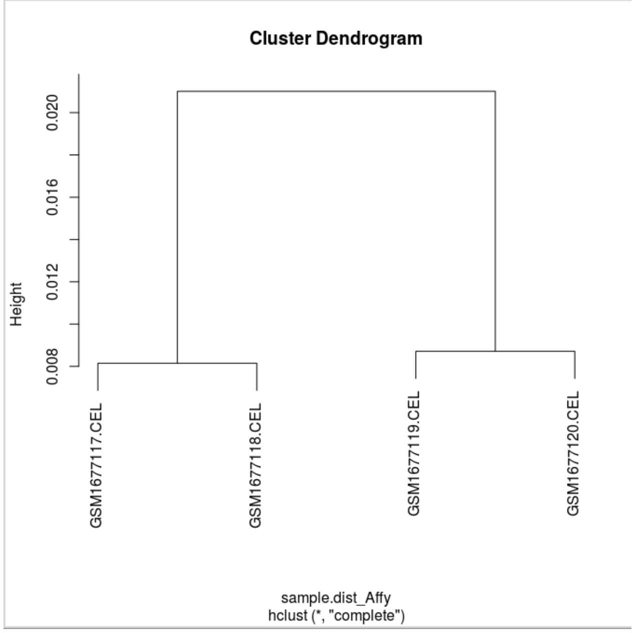
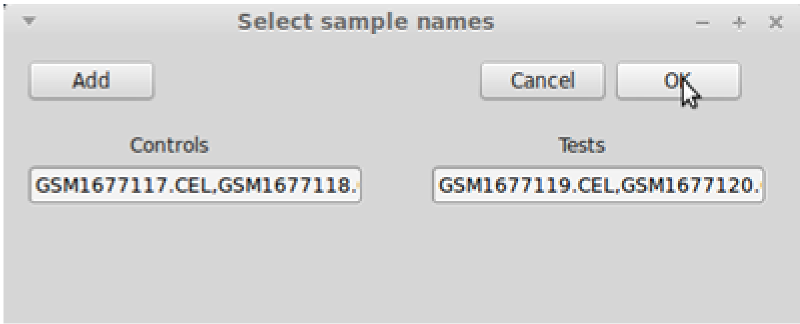


Figure 2.12 (b) Cluster of samples

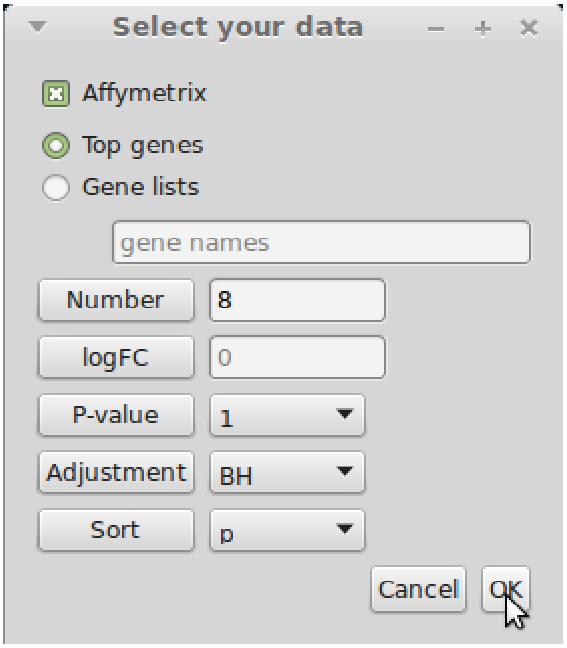
During the microarray data analysis, if the names of controls and tests are not known, unspecific filtering is performed through Analysis → Filtering_and_Statistical_Analysis → UnSpecific. For the present example, control samples are GSM1677117.CEL and GSEM1677118.CEL and the test samples are GSM1677119.CEL and GSEM1677120.CEL. As the control and tests samples are known, specific filtering of samples is performed through Analysis → Filtering_and_Statistical_Analysis → Specific. Control sample names and test sample names are provided in the graphical input box for the specific filtering. It is shown in **Figure 2.12 (c)**. A user can add an extra group for any experiment with more than 1 control or test groups using "Add" button. Once the data is filtered, the top differentially expressed genes are obtained from Analysis → Differential_Gene_Expressions. The parameters available to generate differentially expressed genes are the list of genes names, the number of top differentially expressed genes, log fold change values, p-value cut offs, adjustment methods and the sort options. For the present example, the top eight differentially expressed genes are obtained using the parameters as shown in the following **Figure 2.12 (d)**.



Controls: GSM1677117.CEL,GSM1677118.

Tests: GSM1677119.CEL,GSM1677120.

Figure 2.12 (c) Specific filtering with control and test samples names



☒ Affymetrix

● Top genes

○ Gene lists

Number: 8

logFC: 0

P-value: 1

Adjustment: BH

Sort: p

Figure 2.12 (d) Parameters for differentially expressed genes

Classification of the differentially expressed genes is performed from Analysis → Classification_and_Visualization_Supervized. It is visualized as a heatmap of expression profiles with red color representing up-regulation and green color representing down-regulation of genes as shown in **Figure 2.12 (e)**.

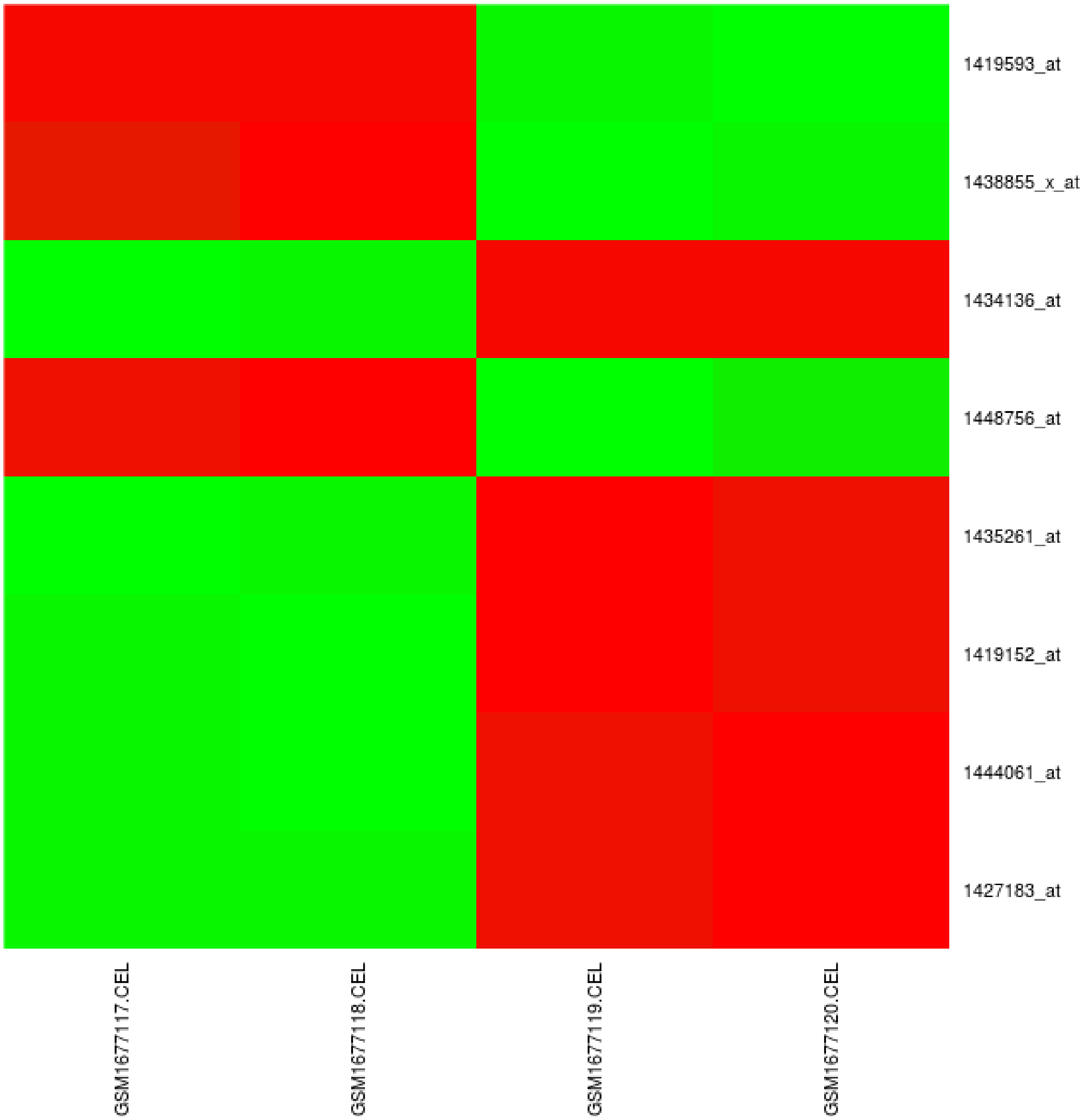


Figure 2.12 (e) Classification of differentially expressed genes

Functional annotations and pathways of the differentially expressed genes are achieved through Miscellaneous → Gene_Set_Enrichment_Analysis. The user can perform annotations and pathways for all the genes in the microarray data through Miscellaneous → Gene_Set_Test_Analysis. Enrichment of differentially expressed genes in GO terms and pathways is visualized as graphs from Miscellaneous → Graphs. In the ontology graphs, yellow colored nodes indicate the significant terms while white nodes are their parents. In the pathway graphs, red colored nodes indicate up-regulation of the genes, green colored nodes indicate down-regulation of the genes and dark grey colored nodes indicate that the genes are not differentially expressed while the white nodes indicate that the genes involved in the pathway are not present in the microarray data. The following **Figure 2.12 (f)** represents the pathway of differentially expressed genes in the KEGG ID “04360”. Node mmu:13176 (light green colored) denoting Dcc gene is down-regulated with a log fold change of 1.5 while the node mmu:19055 (dark green colored) representing Ppp3ca is down-regulated with a log fold change of 3.1.

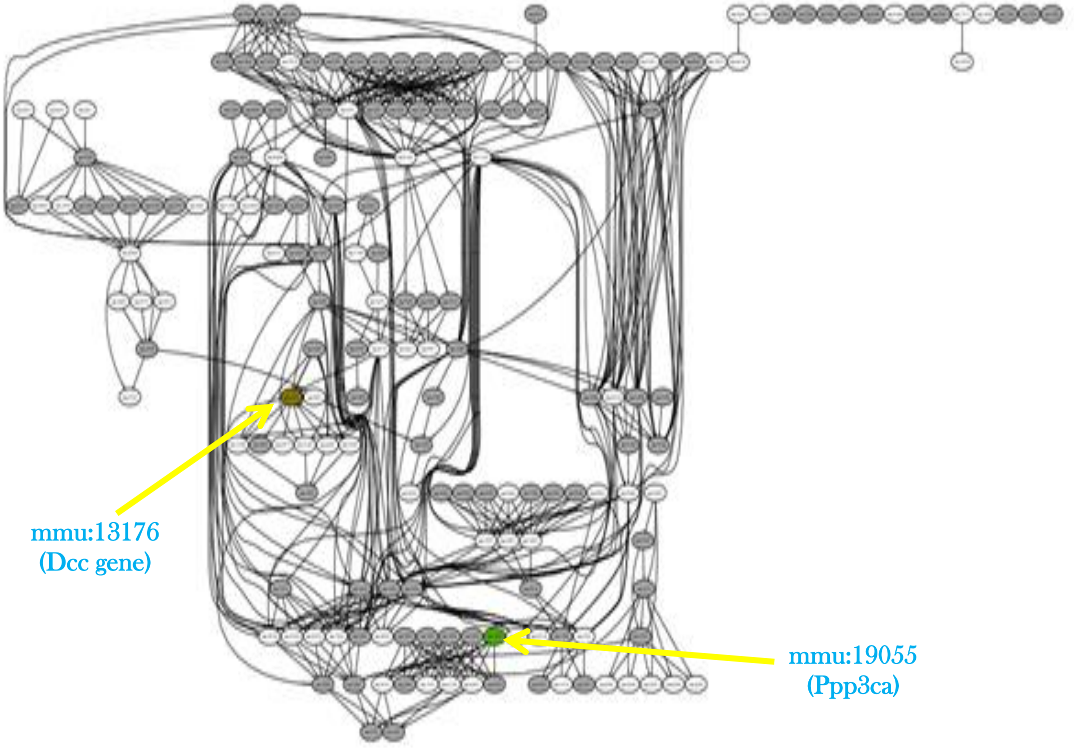


Figure 2.12 (f) Pathway of differentially expressed genes

All the identifiers in the microarray experiment are mapped to corresponding gene symbols from Miscellaneous → Identifier_Symbol. Estimation of sample size for the microarray experiment is achieved through Miscellaneous → Sample_Size_Estimation. Co-expression network is built for differentially expressed genes from Miscellaneous → Coexpression_Network and is shown in the following **Figure 2.12 (g)**.

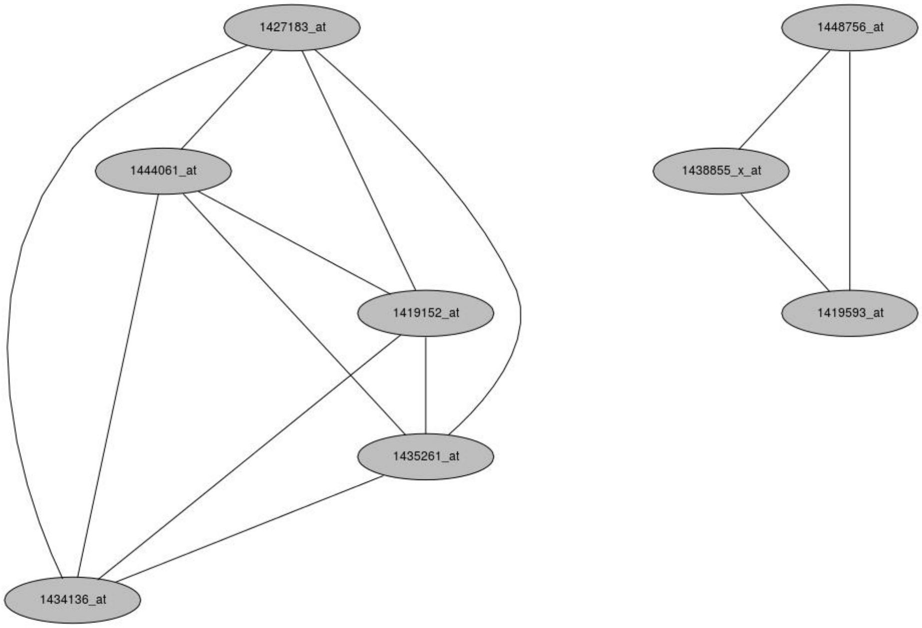
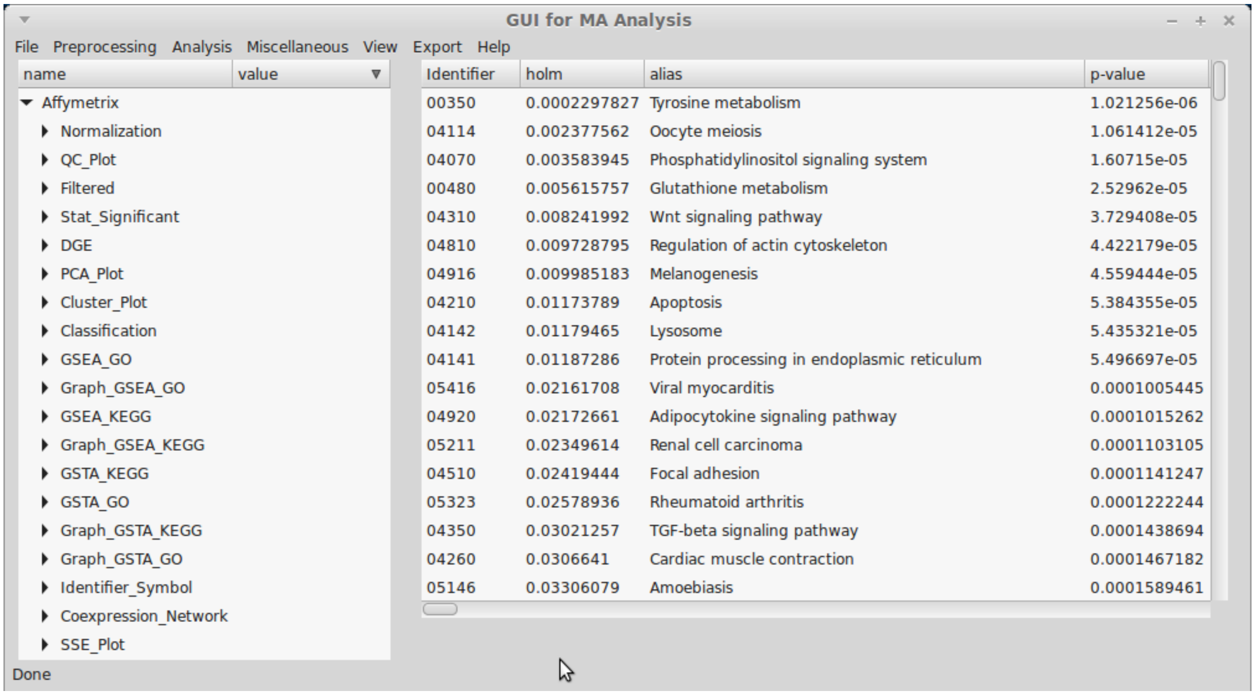


Figure 2.12 (g) Co-expression network of differentially expressed genes

The following **Figure 2.12 (h)** showss the heirarchy of tasks performed during analysis and annotation of microarray data of the present example GSE68613.



name	value
Affymetrix	
Normalization	
QC_Plot	
Filtered	
Stat_Significant	
DGE	
PCA_Plot	
Cluster_Plot	
Classification	
GSEA_GO	
Graph_GSEA_GO	
GSEA_KEGG	
Graph_GSEA_KEGG	
GSTA_KEGG	
GSTA_GO	
Graph_GSTA_KEGG	
Graph_GSTA_GO	
Identifier_Symbol	
Coexpression_Network	
SSE_Plot	

Identifier	holm	alias	p-value
00350	0.0002297827	Tyrosine metabolism	1.021256e-06
04114	0.002377562	Oocyte meiosis	1.061412e-05
04070	0.003583945	Phosphatidylinositol signaling system	1.60715e-05
00480	0.005615757	Glutathione metabolism	2.52962e-05
04310	0.008241992	Wnt signaling pathway	3.729408e-05
04810	0.009728795	Regulation of actin cytoskeleton	4.422179e-05
04916	0.009985183	Melanogenesis	4.559444e-05
04210	0.01173789	Apoptosis	5.384355e-05
04142	0.01179465	Lysosome	5.435321e-05
04141	0.01187286	Protein processing in endoplasmic reticulum	5.496697e-05
05416	0.02161708	Viral myocarditis	0.0001005445
04920	0.02172661	Adipocytokine signaling pathway	0.0001015262
05211	0.02349614	Renal cell carcinoma	0.0001103105
04510	0.02419444	Focal adhesion	0.0001141247
05323	0.02578936	Rheumatoid arthritis	0.0001222244
04350	0.03021257	TGF-beta signaling pathway	0.0001438694
04260	0.0306641	Cardiac muscle contraction	0.0001467182
05146	0.03306079	Amoebiasis	0.0001589461

Done

Figure 2.12 (h) Hierarchy of analysis and annotation of GSE68613

2.4 Conclusions

For the first time, various *in silico* methods are employed in the present study to predict PPIs in *M. tuberculosis*. Sets of PPIs are predicted using *in silico* methods viz., gene neighborhood, phylogenetic tree, mirror tree, gene fusion and interlog. PPIs predicted with multiple *in silico* methods are retrieved. Further, PPIs predicted with any four and five methods are analyzed to understand the the molecular mechanisms of the pathogen. PPIs that are consistently predicted with any two *in silico* methods are filtered through co-expression analysis. Pairs of proteins in the consistently predicted PPIs whose genes are co-expressing in any two microarray experiments are considered confident protein interaction pairs. In total, a set of 8,243 confident PPIs are obtained from the present study. Such interaction data would contribute significantly to achieve a complete protein interactions network and reveal novel signaling pathways as well as potential functions for hypothetical proteins. maGUI, a GUI, is also developed for the analysis of the microarray data. Further, it is used to interpret the genes to the ontology terms and the biological pathways. It could help the user to describe the relationships between genes and characterize specific molecular differences associated with them.

2.5 References

1. Zhou H, Wong L, "Comparative analysis and assessment of *M. tuberculosis* H37Rv protein-protein interaction datasets", BMC Genomics 2011, 12: S20.
2. Valencia A, Pazos F, "Computational methods for the prediction of protein interactions", Curr Opin Struct Biol 2002, 12: 368-373.
3. Eisenberg D, Marcotte EM, Xenarios I, Yeates TO, "Protein function in the post-genomic era", Nature 2000, 405(6788): 823–826.
4. Huynen M, Snel B, Lathe W, Bork P, "Exploitation of gene context", Curr Opin Struct Biol 2000, 10(3): 366–370.

5. Grigoriev A, “A relationship between gene expression and protein interactions on the proteome scale: Analysis of the bacteriophage T7 and the yeast *Saccharomyces cerevisiae*”, Nucleic Acids Res 2001, 29(17): 3513–3519.
6. Ge H, Liu Z, Church GM, Vidal M, “Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*”, Nat Genet 2001, 29(4): 482–486.
7. Jansen R, Greenbaum D, Gerstein M, “Relating whole-genome expression data with protein–protein interactions”, Genome Res 2002, 12(1): 37–46.
8. De Las Rivas J, de Luis A, “Interactome data and databases: different types of protein interaction”, Comp Funct Genomics 2004, 5(2): 173-178.
9. Skrabanek L, Saini HK, Bader GD, Enright AJ, “Computational Prediction of Protein-Protein Interactions”, Mol Biotechnol 2008, 38(1): 1-17.
10. Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA, “Protein interaction maps for complete genomes based on gene fusion events”, Nature 1999, 402(6757): 86-90.
11. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO, “Assigning protein functions by comparative genome analysis: protein phylogenetic profiles”, Proc Natl Acad Sci U S A 1999, 96(8): 4285-4288.
12. Pazos F, Valencia A, “Similarity of phylogenetic trees as indicator of protein-protein interaction”, Protein Eng 2001, 14: 609-614.
13. Matthews LR, Vaglio P, Reboul J, Ge H, Davis BP, Garrels J, Vincent S, Vidal M, “Identification of potential interactions networks using sequence-based searches for conserved protein-protein interactions or interlogs”, Genome Res 2001, 11(12): 2120-2126.
14. Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D, “A combined algorithm for genome wide prediction of protein function”, Nature 1999, 402(6757): 83–86.
15. Govindarajan R, Duraiyan J, Kaliyappan K, Palanisamy M, “Microarray and its applications”, J Pharm Bioallied Sci 2012, 4(Suppl 2): S310-S312.

16. Moreno-Hagelsieb G, Latimer K, “Choosing BLAST options for better detection of orthologs as reciprocal best hits”, Bioinformatics 2008, 24: 319-324.

17. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ, “Basic Local alignment search tool”, J Mol Biol 1990, 215: 403-410.

18. Midha M, Prasad NK, Vindal V, “MycoRRdb: a database of computationally identified regulatory regions within intergenic sequences in mycobacterial genomes”, PLoS One 2012, 7(4): e36094.

19. Tamames J, Casari G, Ouzounis C, Valencia A, “Conserved clusters of functionally related genes in two bacterial genomes”, J Mol Evol 1997, 44(1): 66–73.

20. Dandekar T, Snel B, Huynen M, Bork P, “Conservation of gene order: A fingerprint of proteins that physically interact”, Trends Biochem Sci1998, 23(9): 324–328.

21. Snel B, Bork P, Huynen MA , “Genomes in flux: The evolution of archaeal and proteobacterial gene content”, Genome Res 2002, 12(1): 17–25.

22. Kunin V, Cases I, Enright AJ, de Lorenzo V, Ouzounis CA, “Myriads of protein families, and still counting”, Genome Biol 2003, 4(2): 401.

23. Ouzounis C, Kyrpides N, “The emergence of major cellular processes in evolution”, FEBS Letters 1996, 390(2): 119–123.

24. Rivera MC, Jain R, Moore JE, Lake JA, “Genomic evidence for two functionally distinct gene classes”, Proc Natl Acad Sci U S A 1998, 95(11): 6239–6244.

25. Marcotte EM, Xenarios I, van der Bliek AM, Eisenberg D, “Localizing proteins in the cell from their phylogenetic profiles”, Proc Natl Acad Sci U S A 2000, 97(22): 12115–12120.

26. Pagel P, Wong P, Frishman D, “A domain interaction map based on phylogenetic profiling”, J Mol Biol 2004, 344(5): 1331–1346.

27. Fryxell KJ, “The coevolution of gene family trees”, Trends Genet 1996, 12(9): 364-369.

28. Goh CS, Bogan AA, Joachimiak M, Walther D, Cohen FE, "Co-evolution of proteins with their interaction partners", J Mol Biol 2000, 299(2): 283-293.
29. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins DG, "Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega", Mol Syst Biol 2011, 7: 539.
30. Goh CS, Cohen FE, "Co-evolutionary analysis reveals insights into proteinprotein interactions", J Mol Biol 2002, 324(1): 177-192.
31. Walhout AJM, Sordella R, Lu X, Hartley JL, Temple GF, Brasch MA, Thierry-Mieg N, Vidal M, "Protein interaction mapping in *C. elegans* using proteins involved in vulval development", Science 2000, 287(5450): 116-122.
32. Xenarios I, Salwínski L, Duan XJ, Higney P, Kim SM, Eisenberg D, "DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions", Nucleic Acids Res 2002, 30(1): 303-305.
33. Kerrien S, Aranda B, Breuza L, Bridge A, Broackes-Carter F, Chen C, Duesbury M, Dumousseau M, Feuermann M, Hinz U, Jandrasits C, Jimenez RC, Khadake J, Mahadevan U, Masson P, Pedruzzi I, Pfeiffenberger E, Porras P, Raghunath A, Roechert B, Orchard S, Hermjakob H, "The IntAct molecular interaction database in 2012", Nucleic Acids Res 2012, 40(Database issue): D841-846.
34. Wang Y, Cui T, Zhang C, Yang M, Huang Y, Li W, Zhang L, Gao C, He Y, Li Y, Huang F, Zeng J, Huang C, Yang Q, Tian Y, Zhao C, Chen H, Zhang H, He ZG, "Global Protein-Protein interactions network in the Human Pathogen *Mycobacterium tuberculosis H37Rv",* J Proteome Res 2010, 9(12): 6665-6677.
35. Liu ZP, Wang J, Qiu YQ, Leung RK, Zhang XS, Tsui SK, Chen L, "Inferring a protein interaction map of *Mycobacterium tuberculosis* based on sequences and interologs", BMC Bioinformatics 2012, 13(7): S6.
36. 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, 33: D433-D437.

37. Goll J, Rajagopala SV, Shiau SC, Wu H, Lamb BT, Uetz P, "MPIDB: the microbial protein interaction database", Bioinformatics 2008, 24(15): 1743-1744.
38. Fabregat A, Sidiropoulos K, Garapati P, Gillespie M, Hausmann K, Haw R, Jassal B, Jupe S, Korninger F, McKay S, Matthews L, May B, Milacic M, Rothfels K, Shamovsky V, Webber M, Weiser J, Williams M, Wu G, Stein L, Hermjakob H, D'Eustachio P, "The Reactome pathway Knowledgebase", Nucleic Acids Res 2016, 44(D1): D481-D487.
39. Gentleman R, Huber W, "Making the most of high-throughput protein-interaction data", Genome Biol 2007, 8(10): 112.
40. Von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, "Comparative as-sessment of large-scale data sets of protein-protein interactions", Nature 2002, 417(6887): 399-403.
41. Lee HK, Hsu AK, Sajdak J, Qin J, Pavlidis P, "Coexpression analysis of human genes across many microarray data sets", Genome Res 2004, 14(6): 1085-1094.
42. Qi Y, Bar-Joseph Z, Klein-Seetharaman J, "Evaluation of different biological data and computational classification methods for use in protein interaction prediction", Proteins 2006, 63(3): 490-500.
43. Sprinzak E, Sattath S, Margalit H, "How reliable are experimental protein-protein interaction data?", J Mol Biol 2003, 327(5): 919-923.
44. Tarassov K, Messier V, Landry CR, Radinovic S, Serna Molina MM, Shames I, Malitskaya Y, Vogel J, Bussey H, Michnick SW, "An *in vivo* map of the yeast protein Interactome", Science 2008, 320(5882): 1465-1470.
45. Huang H, Jedynak BM, Bader JS, "Where have all the interactions gone? Estimating the coverage of two-hybrid protein interaction maps", PLoS Comput Biol 2007, 3(11): e214.
46. Hart GT, Ramani AK, Marcotte EM, "How complete are current yeast and human protein-inter-action networks?", Genome Biol 2006, 7(11): 120.

47. Gioffre A, Infante E, Aguilar D, Santangelo MP, Klepp L, Amadio A, Meikle V, Etchechoury I, Romano MI, Cataldi A, Hernandez RP, Bigi F, "Mutation in mce operons attenuates *Mycobacterium tuberculosis* virulence", Microbes Infect 2005, 7(3): 325-334.

48. Forrellad MA, Klepp LI, Gioffre A, Sabio y Garcia J, Morbidoni HR, Santangelo MP, Cataldi AA, Bigi F, "Virulence factors of the *Mycobacterium tuberculosis* complex", Virulence 2013, 4(1): 3-66.

49. Monahan IM, Betts J, Banerjee DK, Butcher PD, "Differential expression of mycobacterial proteins following phagocytosis by macrophages", Microbiology 2001, 147: 459–471.

50. Stewart GR, Wernisch L, Stabler R, Mangan JA, Hinds J, Laing KG, Young DB, Butcher PD, "Dissection of the heat-shock response in *Mycobacterium tuberculosis* using mutants and microarrays", Microbiology 2002, 148: 3129–3138.

51. Dosanjh NS, Rawat M, Chung JH, Av-Gay Y, "Thiol specific oxidative stress response in mycobacteria", FEMS Microbiol Lett 2005, 249(1): 87–94.

52. Rodrigue S, Provvedi R, Jacques PE, Gaudreau L, Manganelli R, "The sigma factors of *Mycobacterium tuberculosis*", FEMS Microbiol Rev 2006, 30: 926–941.

53. Sachdeva P, Misra R, Tyagi AK, Singh Y, "The sigma factors of *Mycobacterium tuberculosis*: regulation of the regulators", FEBS J 2010, 277: 605–626.

54. Lawrence M, Verzani J, "Programming Graphical User Interfaces in R", Chapman and Hall/CRC The R Series 2012.

55. R Core Team, "R: A Language and Environment for Statistical Computing", R Foundation for Statistical Computing, Vienna, Austria 2013.

56. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Lacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini Aj, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, Zhang J, "Bioconductor: Open software

development for computational biology and bioinformatics", Genome Biol 2004, 5(10): R80.

57. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK, "limma powers differential expression analyses for RNA-sequencing and microarray studies", Nucleic Acids Res 2015, 43(7): e47.
58. Dunning MJ, Smith ML, Ritchie ME, Tavare S, "beadarray: R classes and methods for Illumina bead-based data", Bioinformatics 2007, 23(16): 2183-2184.
59. Du P, Kibbe WA, Lin SM, "lumi: a pipeline for processing Illumina microarray", Bioinformatics 2008, 24(13): 1547-1548.
60. Davis S, Meltzer P, "GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor", Bioinformatics 2007, 14: 1846-1847.
61. Zhu Y, Davis S, Stephens R, Meltzer PS, Chen Y, "GEOmetadb: powerful alternative search engine for the Gene Expression Omnibus", Bioinformatics 2008; 24(23): 2798-2800.
62. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G, "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium", Nat Genet 2000, 25(1): 25–29.
63. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K, "KEGG: new perspectives on genomes, pathways, diseases and drugs", Nucleic Acids Res 2017, 45(D1): D353-D361.

CHAPTER 3

Construction and Analysis of Protein Interactions Network of *M. tuberculosis*

3.1 Introduction

The time has arisen for the modern biology to budge from the study of an individual molecule to networked systems. With the advancements in the genomics and proteomics technologies, several genes and proteins are identified in various living organisms. These genes and proteins perform different kinds of molecular functions and participate in various biological processes. Such mechanisms are facilitated by networks of protein-protein interactions (PPIs) and gene regulations. PPIs networks are analyzed to decipher the principal mechanisms and behavior of various biological systems.

The protein interactions network of numerous genomes is poorly understood even for a well-studied organism such as *S. cerevisiae* (yeast) [1]. Enormous data is generated through high throughput methods and the protein relations in the data are of different types which include physical, correlated and collocated interactions [2]. In view of large incompleteness of protein interactions networks and also muddling data scape from various high-throughput methods, a complete protein interactions network is achieved by integrating various types of protein relations [2]. Thus, a complete protein interactions network that operates in a cell is very complicated with a large number of proteins and distinct types of protein relations.

Typically, a PPIs network is represented by an undirected graph with nodes denoting unique proteins and the edges denoting interactions between the proteins. A PPIs network of the data generated from the high-throughput methods is a very large graph with thousands of nodes and tens of thousands of edges. Analysis of such a network is facilitated with the implementation of Graph theory [3]. Graph mining techniques and various biological evidences such as the shared locations, functions and expressions are integrated to reduce search space and time complexity [4]. It not only aid in understanding the inheritance of complex cellular processes at the system-level [5] but also in the development of diagnostics and therapeutics for the diseases.

For *MTB*, many novel features are discovered through genome annotation and comparison studies [6]. Various studies helped in understanding the gene regulation and pathogenesis of the organism [7-8]. Nonetheless, the large number of putative data has made the understanding of the pathogen at the system level inadequate [9-10]. Therefore, systematic study of protein interactions network is required to understand the survival and infection mechanisms of the pathogen. Several attempts are made until recently to predict the protein-protein interactions network of *MTB* [7, 11]. The coverage and consistency of the protein-protein interactions networks are inadequate due to the implementation of a few of the available methods. Therefore, a protein interactions network of *MTB* is built from the confident PPIs generated from the previous *Chapter 2.* It is analyzed to find highly interacting and closely connected proteins. The highly interacting proteins are integrated with different domains of functional annotations to infer biological significance. Further, the protein interactions network is investigated for novel PPIs in *MTB.*

3.2 Materials and Methods

3.2.1 Protein interactions network

Confident PPIs generated from the previous Chapter 2 are used to construct a protein interaction map of *MTB.* This is achieved through the application of VisANT software [12]. Topological properties such as the number of nodes (proteins), number of edges (interactions), degree distribution, clustering coefficient, cliques and hubs are derived from the interactions network. These properties are used to understand the biological significance of the proteins and their interactions and their implications in the current TB research.

3.2.2 Novel protein-protein interactions

Confident PPIs which are consistently predicted with more than two *in silico* methods employed in the present study are obtained. These PPIs are compared with existing literature and databases and the new PPs are identified. Since the methods using sequence information are more universal and steadfast [13-14], the new PPIs are substantiated using machine learning based support vector classification approach and is described in the next section.

3.2.3 Support vector classification

Conjoint triad method [13] is employed to support novel PPIs. It is based on the classification of PPIs using support vector machine. To employ this approach, 20 amino acids are divided into 7 classes based on their similarity in physiochemical properties. The classes are joined into triads to represent a conjoint triad class. Thus any protein can be represented with a total of 343 conjoint triad classes and a protein interaction pair with a total of 686 conjoint triad classes. To prepare positive training examples, interaction pairs of *MTB* retrieved from the DIP database [15] are used. Frequencies of all the classes in an interacting protein from the DIP database are calculated and normalized to generate features of the classes. A positive training example is then represented with classes and features of an interacting protein concatenated with classes and features of the other interacting protein partner. Since the interactions of the proteins are symmetrical, bidirectional positive training examples are generated. Negative training examples are produced from positive training examples by the calculation of pairs of proteins with the highest Euclidean distances values. Both the positive and negative training examples are scaled together using LIBSVM software [16]. The scaled data is randomized 1000 times and used to derive the best C and γ values of RBF (Radial Basis Function) kernel through the grid search approach. Such scaled data along with the best C and γ values are used to train RBF kernel and generate a SVM model. The SVM model is then employed on the novel PPIs to infer the implication of the novel protein interactions.

3.3 Results and Discussion

3.3.1 Protein interactions network

Confident PPIs are used to build the PPIs network of *MTB*. This is achieved using VisANT software. In the interactions network, proteins are represented by green colored circular balls or nodes and the interactions are represented by black colored lines or edges. A single large complex or the core complex is extracted from the interactions network. The following **Figure 3.1** represents the core interactions network of *MTB.*

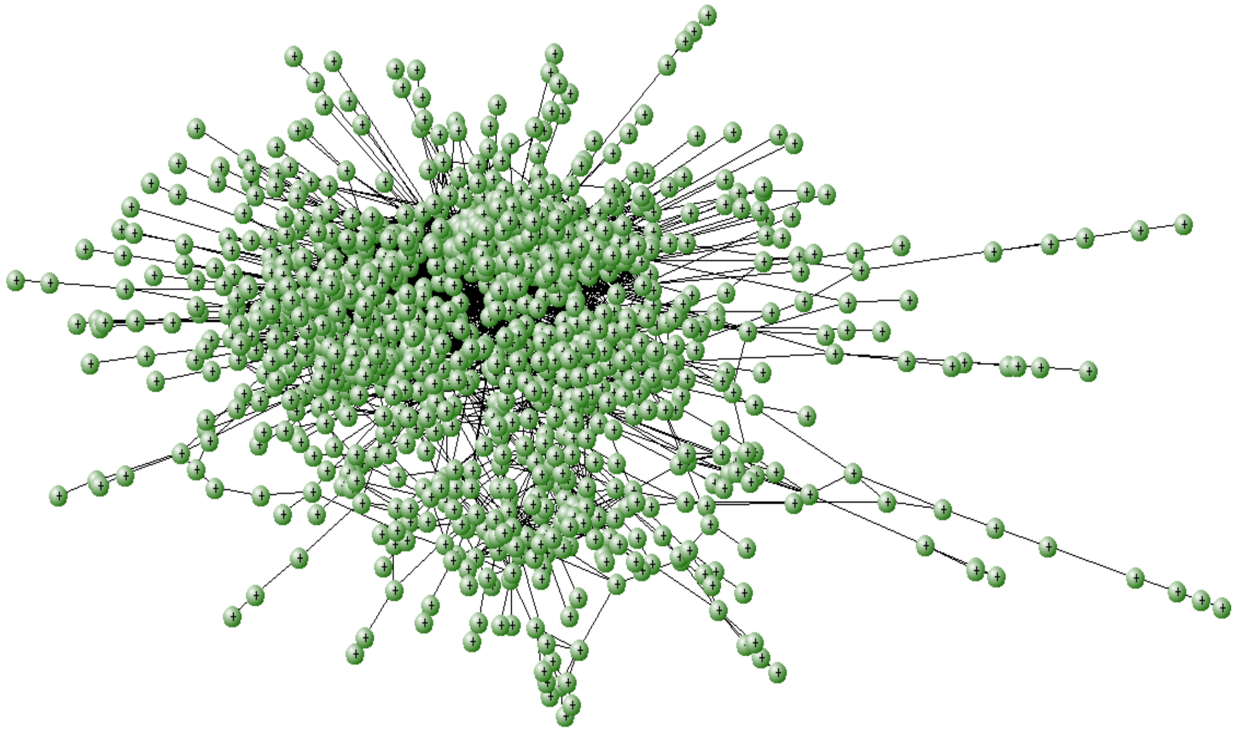


Figure 3.1 A core protein interactions network of *MTB*

It is evident from **Figure 3.1** that it is a simple and undirected network. It is observed that there are 1,086 nodes connected through 8,056 edges in the interactions network. The number of nodes with which a node is connected represents a degree or connectivity [17]. The degree of each node is derived from the interactions network. It is observed from the degree distribution that node degree decreases with an increase in the number of nodes indicating few nodes are connected with large number of nodes while many nodes are connected with only fewer nodes. As the degree distribution obeys the power law, the network is a scale-free network [3, 18]. Further, it is observed that Rv2462c (tig) with a degree of 143, Rv0683 (rpsG) with a degree of 127 and Rv0702 (rplD) with a degree of 124 are the top 3 highest degree nodes in the interactions network. The following **Figure 3.2** represents the degree distribution of the core interaction network.

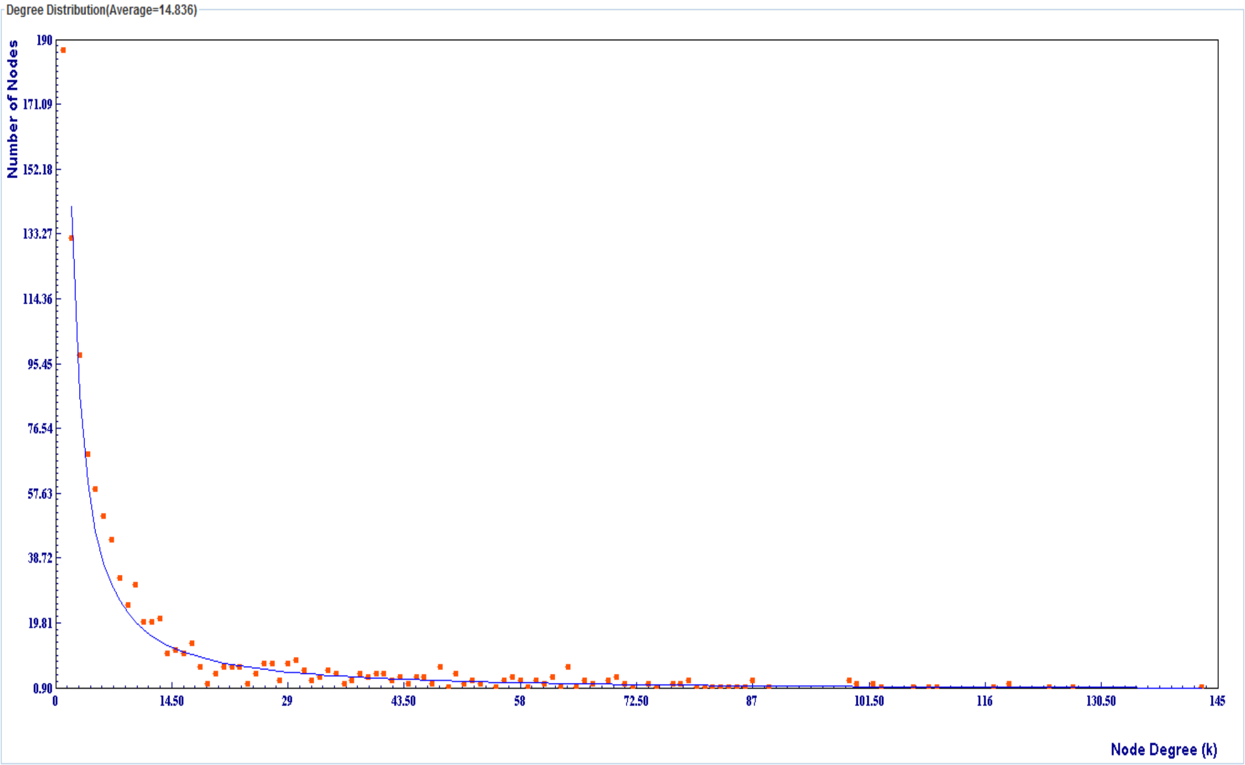


Figure 3.2 Degree distribution of the core interactions network

Clustering coefficient [19-20] of a node specifies the connectivity information about its neighbors. It is used to distinguish highly and sparsely connected structures in a network [20]. It is calculated using the following formula [19]

$$C = \frac{2e}{n(n-1)}$$

where C = clustering coefficient of a node

e = number of edges among the nodes connected to a node

n = number of neighbors of a node

The formula indicates that a node with a large number of interacting partners has lower clustering coefficient values than the nodes with a small number of interacting partners. Correlation of clustering coefficient distribution of the core interactions network is found to be 0.78 with an average of 0.186 indicating that the nodes in the network are finely connected with their neighbors. The following **Figure 3.3** represents the distribution of clustering coefficient values of all the nodes. The figure clearly shows that as the node degree increases, the clustering coefficient value increases.

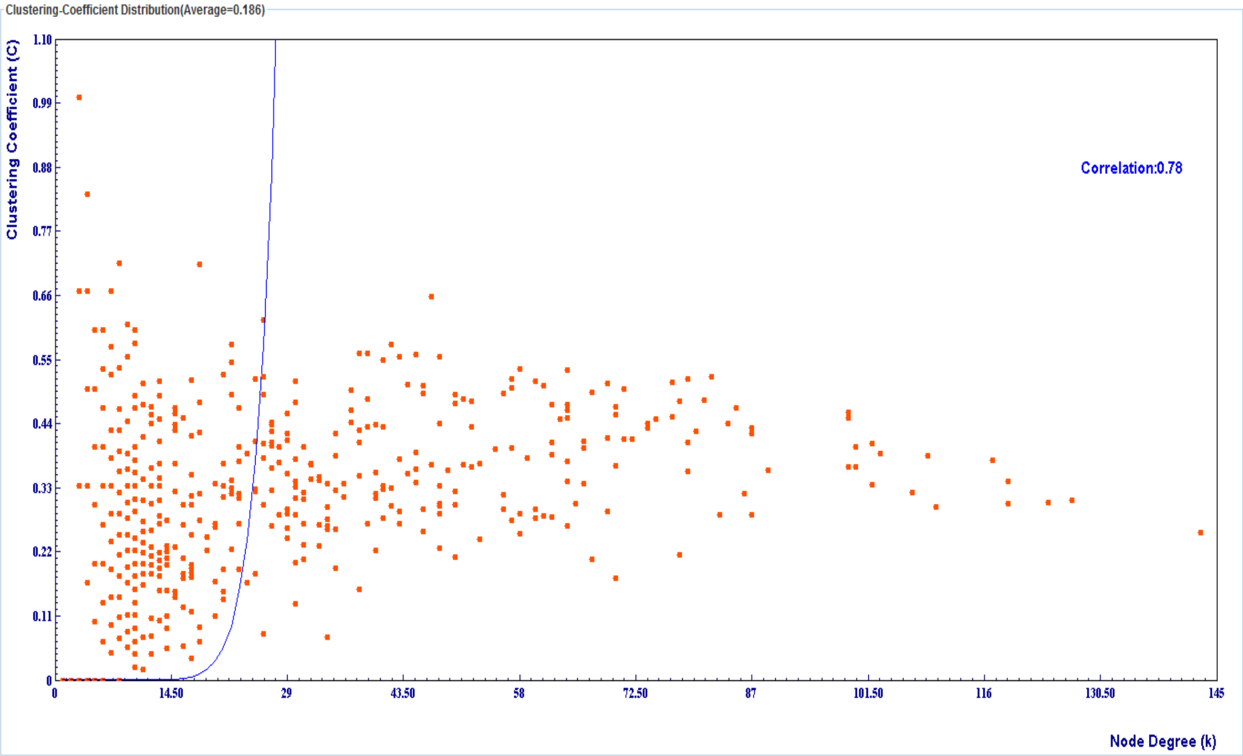


Figure 3.3 Clustering coefficient distribution of the core interactions network

Structures, where every node is connected to every other node in a network, are called cliques [21]. Cliques imply related functions of the nodes. In the present interactions network 3 cliques are observed, viz., Rv2764c (thyA), Rv0501 (galE2) and Rv3329. Therefore, thyA encoding thymidylate synthase, galE2 encoding UDP-glucose 4-epimerase and Rv3329 encoding aminotransferase have related functions and can form a complex protein. Nodes with large number of interacting partners are called hubs. The hubs are known to code for the essential genes [22-23, 17]. They evolve slowly [24-25] and are less susceptible to forfeiture than others during the course of evolution [26-28]. Such nodes are lethal and cause fatality when removed [17, 29]. Hubs are identified in the present interactions network by filtering the nodes with degree cut off of 50 [17]. The number of hub proteins obtained is 95 and are listed in the following **Table 3.1.** The number of interactions with the nodes at the shortest path length of one and two with hubs is 4,839 and 7,201 respectively. Thus deletion of hub proteins causes fatality [17, 29]. Therefore, hub proteins are essential and vital for the integrity and maintenance of the protein interactions network. They can also be potential drug targets [30] against TB disease.

Table 3.1 Set of hub proteins in the core interaction network

adhC	gcvH	lipA	pth	rplE	RpsE	sucC	Rv1697	Rv3193c
Adk	gcvT	lipB	purN	rplF	RpsG	tesB1	Rv1708	Rv3579c
aftB	glcB	lprC	qcrA	rplO	RpsI	Tig	Rv2134c	Rv3677c
aspS	glpD2	murX	recF	rplP	RpsM	trpC	Rv2147c	Rv3678c
atpC	gmk	pgsA2	rfbD	rplQ	RpsP	yajC	Rv2199c	Rv3722c
atpG	greA	pheS	rimM	rplS	RpsQ	Rv0466	Rv2239c	Rv3780
ccsA	gyrB	Ppa	rmlB	rplW	RraA	Rv0546c	Rv2257c	Rv3802c
Csd	hadA	ppiA	rnhB	rplX	ScpA	Rv0556	Rv2509	
dprE2	infC	prcA	rplA	rplY	SecF	Rv1312	Rv2613c	
echA9	lepB	prcB	rplB	rpsC	serB2	Rv1540	Rv2993c	
Frr	leuA	prsA	rplD	rpsD	SseA	Rv1626	Rv3030	

The topological properties of the core interactions network are summarized in the following **Table 3.2.**

Table 3.2 Topological properties of the core interaction

Network Property	Value
Number of Nodes	1,086
Number of Edges	8,056
Network Type	Simple and scale free
Clustering Coefficient correlation	0.78
Number of Cliques	3
Average Degree Distribution	14.836
Top 3 Highest Degree Nodes	143 for tig, 127 for rpsG, 124 for rplD
Hubs	95

3.3.2 Functional annotations of hub proteins

Functional annotations are performed for hub proteins using PANTHER's statistical over-representation test [31]. Ontology terms of hub proteins related to biological processes, molecular functions and cellular component are obtained from the GO database [32] through Fisher's exact test with false discovery rate corrections. For each term, p-value less than 0.05 and more than 3 proteins per function are considered. GO terms related to cellular components of hub proteins have indicated that they are mostly proteasome core complexes, ribosomes and cell wall structures. GO terms related to molecular functions of hubs are found to have mostly hydrolase and RNA binding activity. GO terms related to biological processes of hubs proteins are observed mostly to be involved in the organization of ribonucleoprotein complex subunits, processing of RNAs and growth. The following **Figure 3.4** represents the functional annotations of the hub proteins.

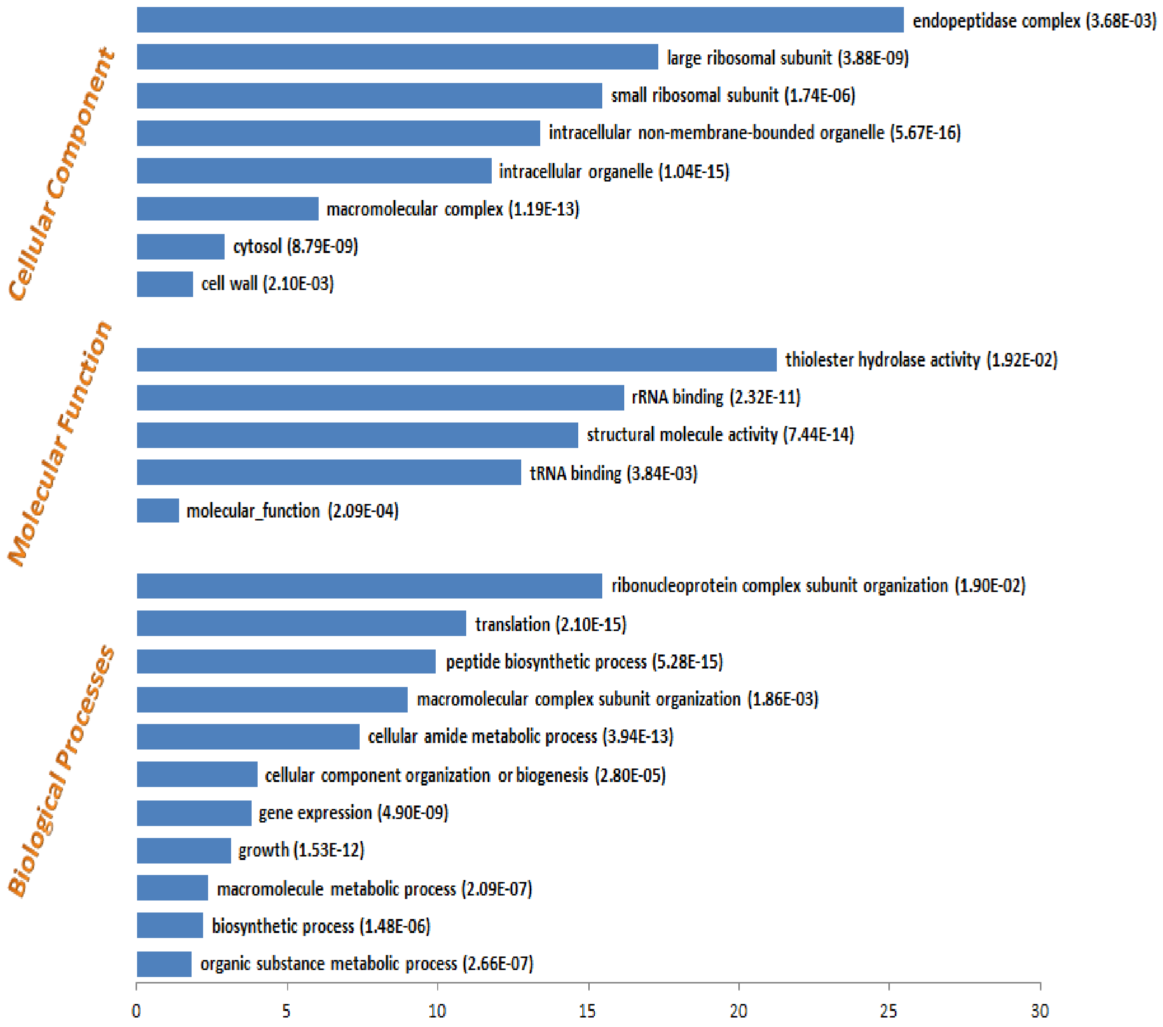


Figure 3.4 Functional annotations of the hub proteins

3.3.3 The novel protein-protein interactions

In the present study, 247 pairs of proteins are found to be interacting with at least three methods and their corresponding genes are observed to be co-expressing in at least two microarray experiments. Of these pairs of proteins, 212 are already known to be interacting in the available literature and databases. The number of overlapping PPIs with the current knowledge is shown in the following **Table 3.3**.

Table 3.3 Overlap of the PPIs in the literature and databases

Sr. No.	Source	Number of Overlapping PPIs in the Source(s)
1	STRING database [33]	164
2	Liu et al data [11]	11
3	Both STRING database and Liu et al data	36
4	All STRING database, Liu et al data, DIP database [15], MPIDB database [34]	1
	Total Overlapping PPIs	**212**

Of the 247 pairs of proteins, 35 are the highly confident interactions pairs which are not reported till date. Therefore, these PPIs are the novel PPIs detected from the present study. The following **Table 3.4** is the list of 35 novel PPIs.

Table 3.4 The novel PPIs

1. ansP2-ansP1
2. csd-rraA
3. dnaJ2-ilvN
4. fadD18-fadD19
5. gmk-hupB
6. greA-rpsI
7. gyrB-lipB
8. gyrB-ppiA
9. hemD-hemB
10. ansP2-ansP1
11. hisC1-hisB
12. mkl-accD5
13. mmaA2-cmaA1
14. mprA-mprB
15. ribH-csd
16. rplD-kdtB
17. rplE-greA
18. rplV-xseA
19. rpsE-kdtB
20. rpsG-greA
21. rpsI-mutY
22. rpsM-mutY
23. tuf-ilvN
24. yrbE1A-yrbE4A
25. aspS-Rv1708
26. aspS-Rv2613c
27. carA-Rv2613c
28. glnQ-Rv0073
29. ppiA-Rv1708
30. prsA-Rv1626
31. regX3-Rv3220c
32. regX3-Rv3579c
33. Rv0068-Rv0439c
34. Rv1378c-Rv3074
35. Rv2165c-Rv2166c

3.3.4 Support vector classification

It is observed that the number of positive and negative examples in the training dataset is 38 and 444 respectively. The grid search of the scaled training data generated the best γ and C values as 0.0078125 and 8.0 respectively. These parameters are used to build a SVM model for the predictions of PPIs. The novel PPIs are tested using the built SVM model. It is observed that the entire novel PPIs fit into the positive training examples. Therefore, the novel PPIs predicted in the present study are the high confident PPIs and it could motivate the current TB research to develop new and improved therapeutic interventions.

3.3.5 Significance of the novel PPIs

The novel PPIs are found to contain 57 proteins. 19 of these are observed to be the hub proteins. **Table 3.5** lists the hub proteins in the novel PPIs.

Table 3.5 List of hubs in the novel PPIs

1. aspS	6. lipB	11. rpsE	16. Rv1626
2. Csd	7. ppiA	12. rpsG	17. Rv1708
3. Gmk	8. prsA	13. rpsI	18. Rv2613c
4. greA	9. rplD	14. rpsM	19. Rv3579c
5. gyrB	10. rplE	15. rraA	

Proteins in the novel PPIs are related with the genes in the DEG database [35]. It is found that 32 proteins in the novel PPIs are the essential gene products. The list of essential genes in the novel PPIs is shown in **Table 3.6.** These genes are vital for survival of the pathogen, hence could be effective drug targets [17, 5].

Table 3.6 List of essential genes in the novel PPIs

1. accD4	9. hemB	17. pks13	25. rpsM
2. accD5	10. hemD	18. prsA	26. Tuf
3. carA	11. hisB	19. rplD	27. Rv1626
4. Csd	12. hisC1	20. rplE	28. Rv1708
5. dnaJ2	13. hupB	21. rplV	29. Rv2165c
6. Gmk	14. ilvN	22. rpsE	30. Rv2166c
7. greA	15. lipB	23. rpsG	31. Rv2613c
8. gyrB	16. mprB	24. rpsI	32. Rv3579c

Proteins of the novel PPIs are found to involve in 7 different pathways as observed from Tuberculist data [36]. The following **Figure 3.5** represents the pathways of the proteins in the novel PPIs. It suggests that 3 proteins are involved in virulence of the pathogen and 4 are conserved hypothetical proteins. Further investigation of these proteins can impact the TB research significantly.

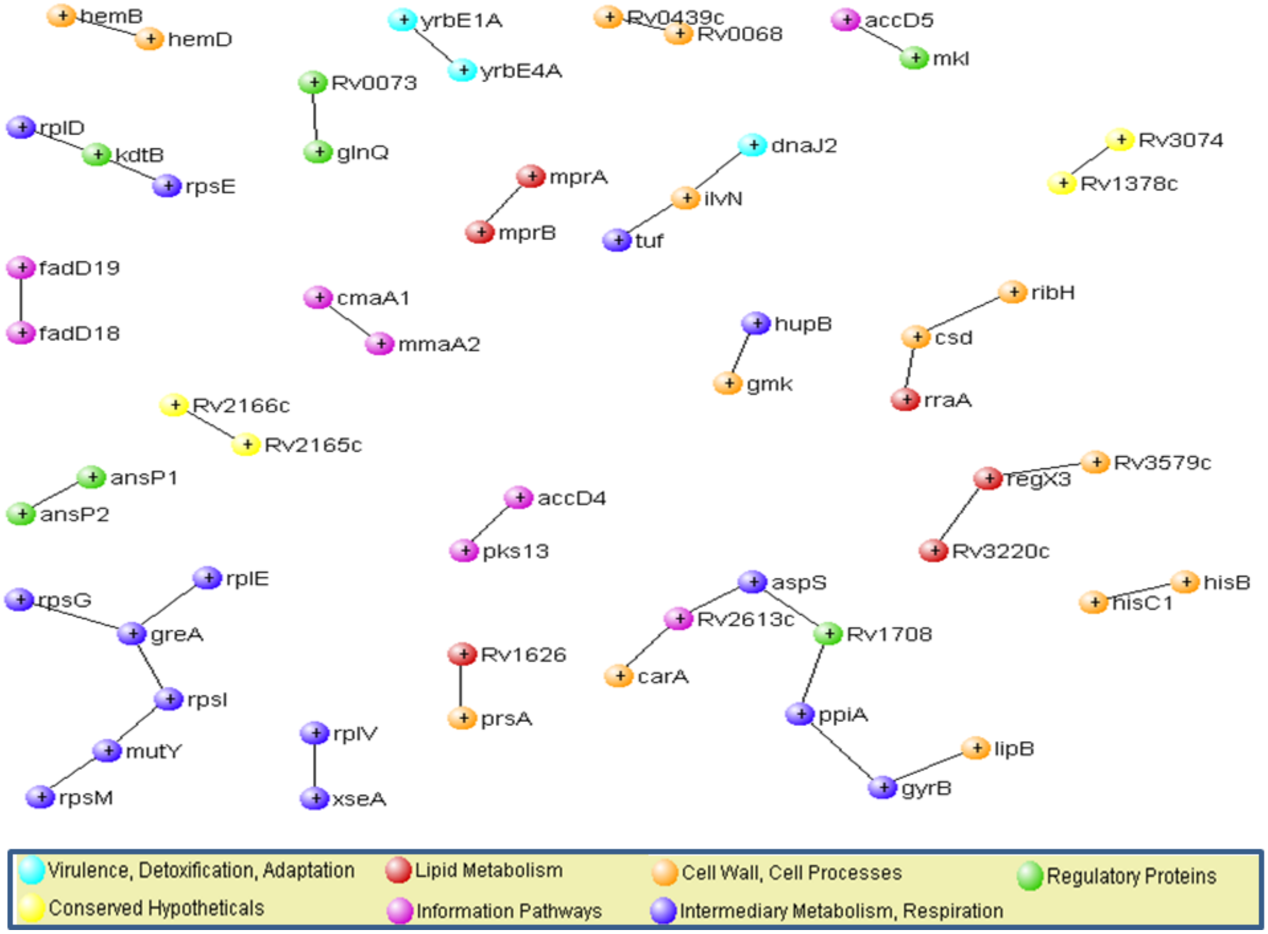


Figure 3.5 Molecular pathways of the proteins in the novel PPIs

3.4 Conclusion

A protein interaction of *MTB* is generated from the confident PPIs. It is used to the core protein interactions network which is found to contain 8,056 interactions connected across 1086 proteins. It is a simple, well clustered and a scale-free network. Highest degree in the protein interactions network is observed for tig protein followed by rspG and rplD proteins. Rv2764c (thyA), Rv0501 (galE2) and Rv3329 proteins are observed to form the completely connected sub-graphs in the interactions network. Using the degree threshold value of 50, a set of 95 hub proteins are obtained. These hubs are mostly proteasome core complexes and involved in cytoplasmic translations and thiolester hydrolase activities. Further, 35 novel PPIs are revealed in the present study. These PPIs are classified as positive interactions through the SVM model built for the prediction of PPIs. The novel PPIs are analyzed to find 19 hub proteins and 32 essential gene products. Functional pathways of the proteins in the novel PPIs are recognized which indicated that yrbE1A, yrbE4A and dnaJ2 are involved in the virulence of the pathogen. Further, high confident interactions among Rv2165c, Rv2166c, Rv1378c and Rv3074 hypothetical proteins are inferred in the present study.

3.5 References

1. Hart GT, Ramani AK, Marcotte EM, “How complete are current yeast and human protein-interactions networks?”, Genome Biol 2006, 7(11): 120.
2. De Las Rivas J, de Luis A, “Interactome data and databases: different types of protein interaction”, Comp Funct Genomics 2004, 5(2): 173-178.
3. Barabasi AL, Oltvai ZN, “Network biology: understanding the cell's functional organization”, Nat Rev Genet 2004, 5(2): 101-113.
4. Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, Emili A, Snyder M, Greenblatt JF, Gerstein M, “A Bayesian networks approach for predicting protein-protein interactions from genomic data”, Science 2003, 302(5644): 449-453.
5. Ideker T, Sharan R, “Protein networks in disease”, Genome Res 2008, 18(4): 644-652.
6. Camus JC, Pryor MJ, Medigue C, Cole ST, “Re-annotation of the genome sequence of *Mycobacterium tuberculosis H37Rv*”, Microbiology 2002, 148(Pt 10): 2967-2973.
7. Wang Y, Cui T, Zhang C, Yang M, Huang Y, Li W, Zhang L, Gao C, He Y, Li Y, Huang F, Zeng J, Huang C, Yang Q, Tian Y, Zhao C, Chen H, Zhang H, He ZG, “Global Protein-Protein interactions network in the Human

Pathogen *Mycobacterium tuberculosis H37Rv*", J Proteome Res 2010, 9(12): 6665-6677.

8. Rapanoel HA, Mazandu GK and Mulder NJ, "Predicting and analyzing interactions between *Mycobacterium tuberculosis* and its human host", PLOS One 2013, 8: e67472.
9. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV, Eiglmeier K, Gas S, Barry CE 3rd, Tekaia F, Badcock K, Basham D, Brown D, Chillingworth T, Connor R, Davies R, Devlin K, Feltwell T, Gentles S, Hamlin N, Holroyd S, Hornsby T, Jagels K, Krogh A, McLean J, Moule S, Murphy L, Oliver K, Osborne J, Quail MA, Rajandream MA, Rogers J, Rutter S, Seeger K, Skelton J, Squares R, Squares S, Sulston JE, Taylor K, Whitehead S, Barrell BG, "Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence", Nature 1998, 393(6685): 537-544.
10. Goulding CW, Parseghian A, Sawaya MR, Cascio D, Apostol MI, Gennaro ML, Eisenberg D, "Crystal structure of a major secreted protein of *Mycobacterium tuberculosis*-MPT63 at 1.5-A resolution", Protein Sci 2002, 11(12): 2887-2893.
11. Liu ZP, Wang J, Qiu YQ, Leung RK, Zhang XS, Tsui SK, Chen L, "Inferring a protein interaction map of *Mycobacterium tuberculosis* based on sequences and interlogs", BMC Bioinformatics 2012, 13(7): S6.
12. Hu Z, Mellor J, Wu J, DeLisi C, "VisANT: an online visualization and analysis tool for biological interaction data", BMC Bioinformatics 2004, 5: 17.
13. Shen J, Zhang J, Luo X, Zhu W, Yu K, Chen K, Li Y, Jiang H, "Predicting protein–protein interactions based only on sequences information", PNAS 2007, 104(11): 4337-4341.
14. Bharne D, Naresh D, Vindal V, "Inferring protein interactions network of *Mycobacterium tuberculosis H37Rv* using sequence information", Res J Life Sci Bioinform Pharm Chem Sci 2018, 4(6): 57-64.

15. Xenarios I, Salwínski L, Duan XJ, Higney P, Kim SM, Eisenberg D, “DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions”, Nucleic Acids Res 2002, 30(1): 303-305.
16. Chang CC, Lin CJ, “LIBSVM: a library for support vector machines”, ACM Trans Intell Syst Technol 2011, 2(3): 2:27:1–2:27:27.
17. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL, “Lethality and centrality in protein networks”, Nature 2001, 411: 41-42.
18. Barabasi AL, Albert R, “Emergence of scaling in random networks”, Science 1999, 286: 509-512.
19. Watt DJ, “Small worlds”, Princeton University Press 1999.
20. Yu H, Greenbaum D, Xin H, Lu XZ, Gerstein M, “Genomic analysis of essentiality within protein networks”, Trends Genet 2004, 20(6): 227-231.
21. Watts DJ, Strogatz SH, “Collective dynamics of ‘small-world’ networks”, Nature 1998, 393(6684): 440-442.
22. He X, Zhang J, “Why do hubs tend to be essential in protein networks?”, PLoS Genet 2006, 2(6): e88.
23. Yu H, Greenbaum D, Xin Lu H, Zhu X, Gerstein M, “Genomic analysis of essentiality within protein networks”, Trends Genet 2004, 20(6): 227-231.
24. Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW, “Evolutionary rate in the protein interactions network”, Science 2002, 296(5568): 750-752.
25. Fraser HB, Wall DP, Hirsh AE, “A simple dependence between protein evolution rate and the number of protein-protein interactions”, BMC Evol Biol 2003, 3(1): 11.
26. Krylov DM, Wolf YI, Rogozin IB, Koonin EV, “Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution”, Genome Res 2003, 13(10): 2229-2235.
27. Wuchty S, “Evolution and topology in the yeast protein interactions network”, Genome Res 2004, 14(7): 1310-1314.
28. Wuchty S, Barabasi AL, Ferdig MT, “Stable evolutionary signal in a yeast protein interactions network”, BMC Evol Biol 2006, 6: 8.

29. Raman K, Vashisht R, Chandra N, “Strategies for efficient disruption of metabolism in *Mycobacterium tuberculosis* from network analysis”, Mol Biosyst 2009, 5(12): 1740-1751.

30. Hase T, Tanaka H, Suzuki Y, Nakagawa S, Kitano H, “Structure of protein interactions networks and their implications on drug design”, PLoS Comput Biol 2009, 5(10): e1000550.

31. Mi H, Poudel S, Muruganujan A, Casagrande JT, Thomas PD, “PANTHER version 10: expanded protein families and functions, and analysis tools”, Nucleic Acids Res 2016, 44(D1): D336-D342.

32. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G, “Gene ontology: tool for the unification of biology. The Gene Ontology Consortium”, Nat Genet 2000, 25(1): 25–29.

33. 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, 33: D433-D437.

34. Goll J, Rajagopala SV, Shiau SC, Wu H, Lamb BT, Uetz P, “MPIDB: the microbial protein interaction database”, Bioinformatics 2008, 24(15): 1743-1744.

35. Luo H, Lin Y, Gao F, Zhang CT, Zhang R, “DEG 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements”, Nucleic Acids Res 2014, 42(Database issue): D574-D580.

36. Lew JM, Mao C, Shukla M, Warren A, Will R, Kuznetsov D, Xenarios I, Robertson BD, Gordon SV, Schnappinger D, Cole ST, Sobral B, “Database resources for the TB community”, Tuberculosis (Edinb) 2013, 93(1): 12-17.

CHAPTER 4

Identification of *Human* and *M. tuberculosis* Protein Interactions using Computational methods

4.1 Introduction

TB is a major contagious disease stated to cause more than a million deaths in 2016 (WHO Global TB Report, 2017). With the increasing TB cases and the rising of multi-drug resistant strains (WHO Report on Multidrug and extensively drug-resistant TB (M/XDR-TB), 2010), it has necessitated for the better understanding of *MTB*. The pathogen infects the human cell through direct contact with the surface proteins [1]. It evades the human immune response and inhibits the macrophage cells. Surviving inside the macrophage cells, it makes several molecular interactions and acquires nutrients for its growth [2]. Understanding these molecular interactions between the human host and the pathogen provides new insights into the infection, survival and persistent system of the pathogen.

The current high-throughput experimental techniques are operated to detect protein interactions at the genomic level within a single organism [3-8]. A few experimental techniques are available for the detection of the host pathogen protein interactions. The number of host-pathogen protein interactions generated from these techniques is inadequate [9]. Resources for the retrieval of host-pathogen protein interactions are also limited [10]. Moreover, the data related to the simultaneous expression of both host and the pathogen during infection hardly existed. Therefore, the elucidation of the host-pathogen protein interactions using experimental techniques is not feasible [10]. In this regard, the current studies depend on the computational methods to detect and interpret the host-pathogen protein interactions. Regrettably, these methods too are developed for the prediction of intra-species protein interactions [11-13]. Hence, a systematic method for the detection and interpretation of cross-species protein interactions is lacking.

Recently, the interactions among the proteins of *Homo sapiens (Human)* and *MTB* are detected and evaluated using the Interlog method [14-15]. During infection, the pathogen makes direct contact with the *Human* cells. Therefore, the methods that detect physical interactions among the *Human* and *MTB* proteins are more crucial. *In silico* Two Hybrid system (I2H) is one of the methods that detect proteins physical interactions within a single species [16]. It exploits the concept of correlated mutations [17] to detect PPIs. It is more likely that the *Human* proteins and the *MTB* proteins undergo correlated mutations during the course of evolution. Therefore, in the present study, I2H is employed to predict physical interactions among the proteins of *MTB* and the *Human* host (*Human-M. tuberculosis* protein interactions or *HMIs*). The Interlog method is employed to detect consistent *HMIs*. Further, functional annotations and co-expression analysis are implemented to generate confident *HMIs*. Such a study supports the current research on TB through the improved understanding of the pathogen infection and survival strategies.

4.2 Material and Methods

4.2.1 Prediction of *Human* and *M. tuberculosis* protein interactions

Physical interactions between the *Human* and *MTB* proteins are predicted using I2H [16] in the present study. Further, the *HMIs* are predicted using Interlog method [14]. The following computational methods describes the I2H and Interlog methods to predict *HMIs*.

4.2.1.1 In silico Two Hybrid system

I2H is based on the concept that two physically interacting proteins undergo mutations at the same time. Such correlated mutations are quantified in the co-evolving genomes [18]. For better coverage of species in the evolutionary range of *MTB* to *Human*, seven model organisms are selected, viz., *E. coli, Dictyostelium discoideum, S. cerevisiae, C. elegans, D. melanogaster, Danio rerio, and M. musculus*. Total proteome sets of *Human*, *MTB* and the model species are retrieved from NCBI genome (https://www.ncbi.nlm.nih.gov/genome/). RBBH [19-20], as discussed in the Chapter 2, is performed on the proteome sets of model species with the proteome set of *Human* to generate *Human* orthologous proteins table. With the similar procedure, *MTB* orthologous proteins table is generated. A multiple sequence alignment (MSA) of a *Human* protein with its orthologs is generated using ClustalO software [21]. Similarly, a MSA is generated for a *MTB* protein with its orthologs. If the number of coincident species in both the MSAs is above 4, the MSAs of the *Human* and *MTB* proteins are reduced to coincident species and compared. A matrix is created for every position in the MSAs using all the unique residues at that position. The matrix is filled with values from McLachlan distance matrix [22]. The matrix of every position in a MSA is compared with the matrix of every other position in the same MSA and an intra-protein correlation values distribution is obtained. Thus, intra-protein correlation values distribution of both *Human* and *MTB* are obtained. The matrix of every position in the *Human* MSA is compared with the matrix of every position in the *MTB* MSA and an inter-protein correlation values distribution of *Human* and *MTB* is obtained. The intra-protein and inter-protein correlation values above 0.4 are used to measure interaction index value using the following formula [16]:

$$\boxed{I_{Mtb\_Hu} = \sum_{i=0.4}^{1} \frac{C(_{Mtb\_Hu})_i}{C(_{Mtb\_Mtb})_i + C(_{Hu\_Hu})_i}\, i}$$

where, I_{Mtb_Hu} = Interaction index value

$C(_{Mtb\_Hu})$ = Inter-protein correlation value

$C(_{Mtb\_Mtb})$ = Intra-protein correlation value of *MTB*

$C(_{Hu\_Hu})$ = Inter-protein correlation values of *Human*

i = correlation value bin (>=0.4)

If the interaction index value is greater than or equal to 2, the *Human* and *MTB* proteins are strongly predicted as the interaction pair [16]. The following **Figure 4.1** represents the systematic procedure for the implementation of I2H to predict *HMIs*.

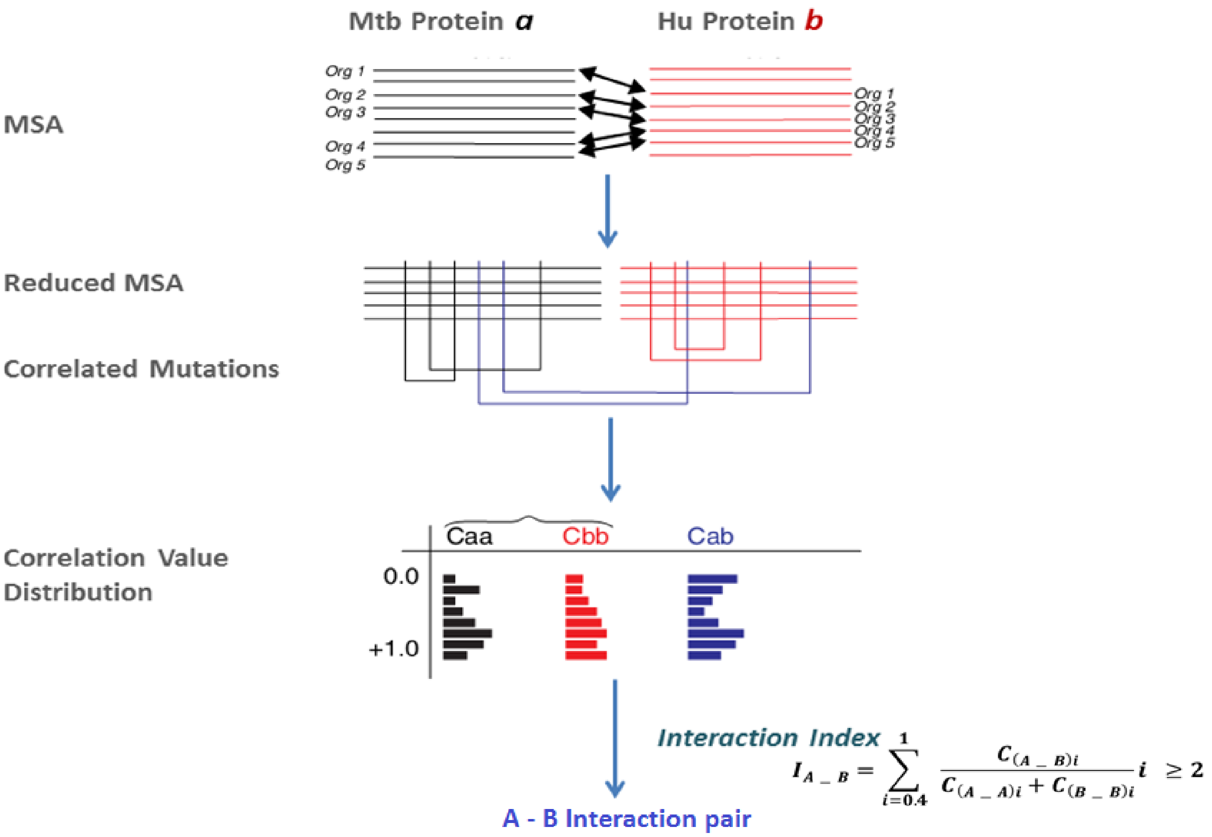


Figure 4.1 *In silico* two hybrid system

4.2.1.2 Interlog method

Experimentally predicted PPIs are retrieved from the DIP database [23]. Physically interacting PPIs are retrieved from the HitPredict database [24]. Experimental, physical and co-expression based PPIs are retrieved from the STRING database [25]. Genes related to *Human* and *MTB* are filtered out from the co-expression based PPIs to avoid redundancy. RBBH is performed on the protein sequences of the retrieved PPIs with the protein sequences of *MTB* to generate a *MTB* orthologous proteins table. Similarly, RBBH is performed with protein sequences of *Human* to generate a *Human* orthologous proteins table. The *MTB* orthologous proteins table and the *Human* orthologous table are compared with the retrieved PPIs. If a *Human* protein is orthologous to one of the proteins in the retrieved PPIs and a *MTB* protein is orthologous to another interaction partner of the same PPI, the *Human* and *MTB* proteins are taken as an interacting protein pair [14, 26]. The following **Figure 4.2** shows the Interlog method for *HMIs* prediction.

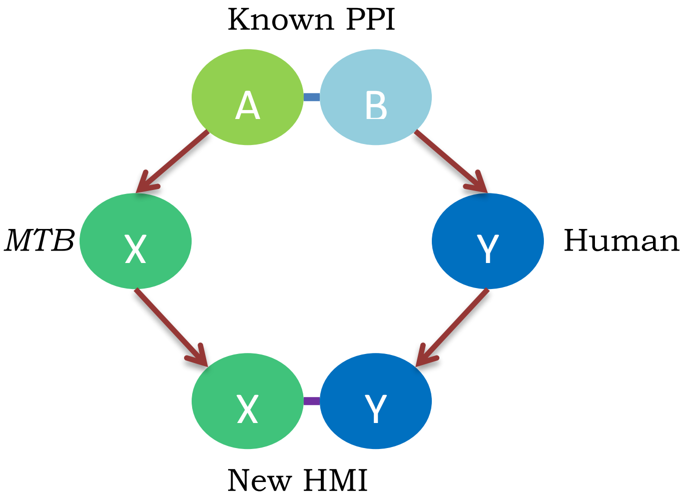


Figure 4.2 Interlog method

4.2.2 Confident *Human* and *M. tuberculosis* protein interactions

The consistent *HMIs* are obtained with the consideration of *HMIs* predicted by both I2H and Interlog methods. Currently, *HMIs* in the literature and databases are very scarce [8-9]. Therefore, functional annotations and co-expression analysis is performed on the consistent *HMIs* to generate confident *HMIs*.

4.2.2.1 Functional annotations

To perform functions annotations, first the *Human* proteins and the *MTB* proteins are separated from the consistent *HMIs*. Using PANTHER software from the Gene Ontology Consortium [27-28], the functional annotation is performed on the *Human* proteins. PANTHER statistical over-representation test is employed on the *Human* proteins using Fisher's exact test type and false discovery rate corrections. The test fits the proteins into different ontology terms related to biological processes, molecular functions and cellular components. The ontology terms of the *Human* proteins are filtered with the p-value less than 0.05 and minimum 2 fold enrichment with at least 3 proteins per function. The similar procedure is employed on the *MTB* proteins and ontology terms are obtained. Ontology terms of the *Human* proteins and *MTB* proteins are compared. If a *Human* protein and a *MTB* have the same ontology term, the corresponding consistent *HMI* is favored as a confident interaction [29].

4.2.2.2 Co-expression analysis

To perform co-expression analysis, pairs of *Human* proteins with common *MTB* interaction partners are extracted from the consistent *HMIs*. List of *Human* proteins are separated from the extracted *HMIs*. These *Human* proteins are used for the co-expression analysis. In this regard, the NCBI GEO microarray experiments [30] are filtered with "Tuberculosis", "*Homo sapiens*", "Expression profiles" and the sample size with at least 10 [31-32]. Further, the experiments are filtered with Bioconductor annotation packages [33] to avoid gene mapping errors. Such experimental data is preprocessed by log base 2 transformation and the k-nearest neighbor imputation and then normalized. The normalized data is filtered for those genes which encode the *Human* proteins in the list for the co-expression analysis. The expression profiles of these genes are used to calculate correlation coefficient values. Pairs of genes with the correlation coefficient value of at least 0.8 in multiple microarray experiments are considered to be co-expressing genes [32]. The co-expressing *Human* genes are compared with consistent *HMIs*. Interactions of the co-expressing *Human* genes products with the common *MTB* interaction partner are favored as confident *HMIs* [29]. The similar procedure is employed for the pairs of *MTB* proteins with the common *Human* interaction partners. List of *MTB* proteins are separated from these pairs and used for the co-expression analysis. The microarray experiments are filtered with "*Mycobacterium tuberculosis*" and a sample size of at least 10 [32]. The experimental data which comprise ORF names is preprocessed and normalized as described above for the *Human* proteins. The normalized data is filtered for the genes which encode the *MTB* proteins in the list for the co-expression analysis. Co-expressing *MTB* genes are obtained by employing the similar procedure described for the

Human proteins. Interactions of the co-expressing *MTB* genes products with the common *Human* interaction partner in the consistent *HMIs* are favored as confident interactions [29]. The following **Figure 4.3** is the flow chart for obtained confident *HMIs*.

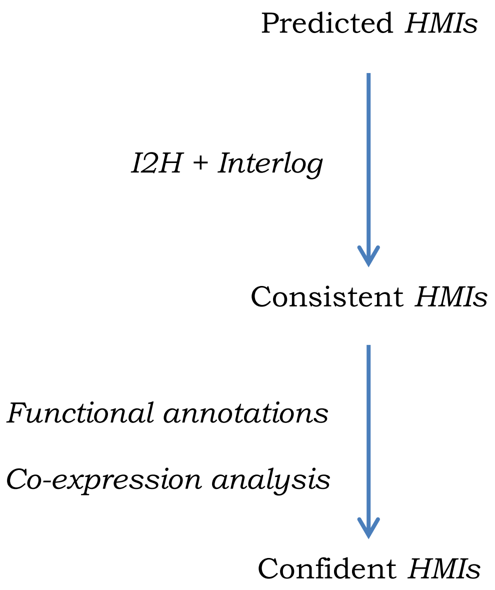


Figure 4.3 Flow-chart for the confident *HMIs*

4.3 Results and Discussion

4.3.1 Predicted *Human* and *M. tuberculosis* protein Interactions

Implementation of the I2H as discussed in *Section 4.2.1.1*, has predicted 20,891 *HMIs*. Using the Interlog method as discussed in *Section 4.2.1.2*, 73 *HMIs* are predicted from DIP data, 140 *HMIs* predicted from HitPredict data and 7,711 *HMIs* predicted from STRING data. It is observed that some *HMIs* are commonly predicted from these multiple data. Therefore, of the total *HMIs* predicted using Interlog methods, 7,788 are the unique *HMIs* obtained.

4.3.2 Significant *HMIs*

The predicted *HMIs* are assessed in the literature and databases which are currently scarcely exist for the protein interactions related to *M. tuberculosis* with *Human*. Nonetheless, 3 *HMIs* are found in a recent report of *Rapanoel* et al [15] and is shown in the following **Table 4.1**. Novel strategies are required to be invented and investigated for the detection of host pathogen protein interactions especially related to the *Human* and *MTB* proteins.

Table 4.1 *HMIs* overlapping in the literature

Sr. No.	Mtb protein partner	Human protein partner
1	atpA	ATP5B
2	atpD	ATP5A1
3	scoB	OXCT1

4.3.3 Consistent *HMIs*

HMIs obtained from I2H are compared with the *HMIs* obtained through Interlog methods from DIP, STRING and HitPredict data. It is observed that 448 *HMIs* are predicted by I2H and Interlog from the DIP data, 22 *HMIs* are predicted by I2H and Interlog from the STRING data and 56 *HMIs* are predicted by I2H and Interlog from the HitPredict data. The following **Figure 4.4** represents the *HMIs* predicted by both I2H and Interlog methods. The total number of unique *HMIs* consistently predicted by both the methods is 485.

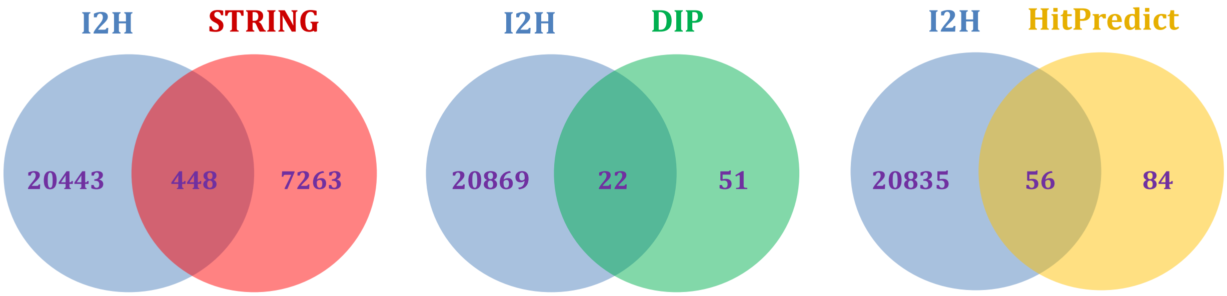


Figure 4.4 Consistent *HMIs*

4.3.4 Confident *HMIs*

Confident *HMIs* are generated through functional annotations of *Human* and *MTB* proteins and co-expression analysis of genes of *Human* proteins and *MTB* proteins as discussed in the *Section 4.2.2*.

4.3.4.1 Functional annotations

It is observed from the functional annotations of Human proteins that they mostly are the components of plastids, perform DNA directed RNA polymerase activity and involve in processes related to tricarboxylic acid cycle and protein folding. It is observed from the functional annotations of *MTB* proteins that they mostly are the segments of proton-transport ATP synthase complexes, perform functions related to ligand-gated ion channel activity, and involve in the generation of precursor metabolite and energy processes. Further, 4 ontology terms related to biological processes, 3 related to molecular functions and 2 related to cellular components are found to be common for both Human and *MTB* proteins and is shown in **Table 4.2.** *Human* and *MTB* proteins of the common ontology terms are combined *[29]* to generate 8,802 *HMIs*. Comparing these *HMIs* with the consistent *HMIs*, 283 confident *HMIs* are obtained.

Table 4.2 Common functional annotations of *Human* and *MTB proteins*

Sr. No.	**Biological Processes Terms**	***Human***	***MTB***	***HMIs***
1	generation of precursor metabolites and energy	25	4	100
2	nucleobase-containing compound metabolic process	110	7	770
3	primary metabolic process	168	12	2,016
4	metabolic process	189	14	2,646
	Molecular Function Terms			
1	proton-transporting ATP synthase activity, rotational mechanism	3	2	6
2	hydrogen ion transmembrane transporter activity	7	2	14
3	catalytic activity	130	14	1,820
	Cellular Component Terms			
1	protein complex	66	3	198
2	Intracellular	154	8	1,232
	Total *HMIs*			8,802
	Overlap with Consistent *HMIs*			**283**

4.3.4.2 Co-expression analysis

In the confident *HMIs*, 22 *MTB* proteins are found to interact with minimum one common *Human* protein. Therefore, co-expression analysis of these *MTB* proteins is performed. In order to achieve this, the microarray experiments are retrieved from NCBI GEO. With the filtering of the experiments with a sample size greater than or equal to 10 [32], 75 experiments are obtained. Further, only 57 of these experiments are found to contain ORF names. Such experiments are used for co-expression analysis as discussed in *Section 4.2.2.2.* Using this approach, a set of 177 confident *HMIs* are obtained. The following **Figure 4.5** is a systematic procedure for generation of confident *HMIs* using *MTB* microarray data.

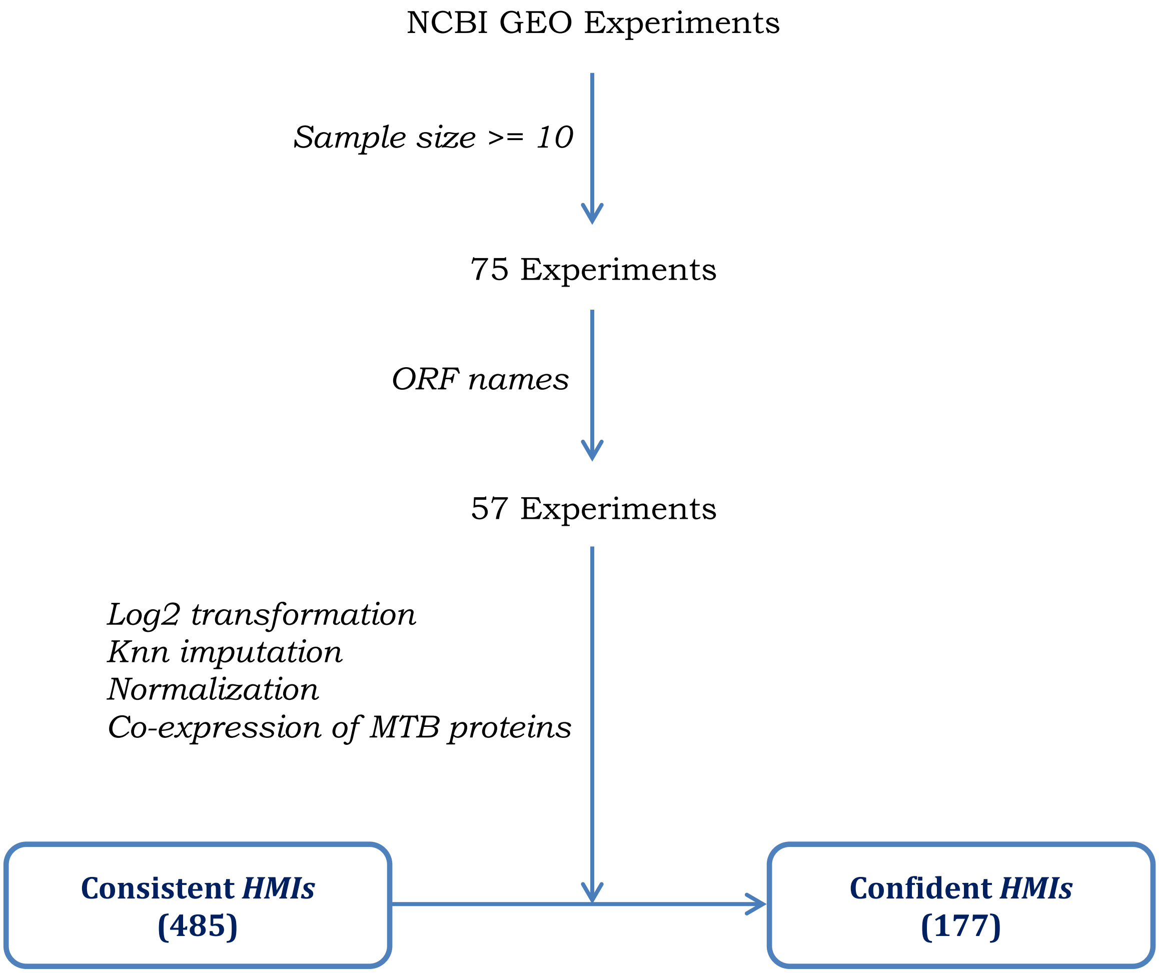


Figure 4.5 Procedure for co-expression analysis of *MTB genes*

It is observed that 290 *Human* proteins are found to interact with minimum one common *MTB* protein. Therefore, co-expression analysis of these Human proteins is performed. The experimental data related to "Tuberculosis", "*Homo sapiens*" and "Expression profiles" *[31]* is retrieved from NCBI GEO. With filtering of the experiments with a sample of 10 and above *[32]*, 43 experiments are generated. These experiments are further filtered with Bioconductor annotation packages *[33]* and 35 experiments are obtained. Co-expression analysis of these experiments as discussed in *Section 4.2.2.2* is performed to detect 230 confident *HMIs*. The following **Figure 4.6** is a systematic procedure for generation of confident *HMIs* from Human microarray data.

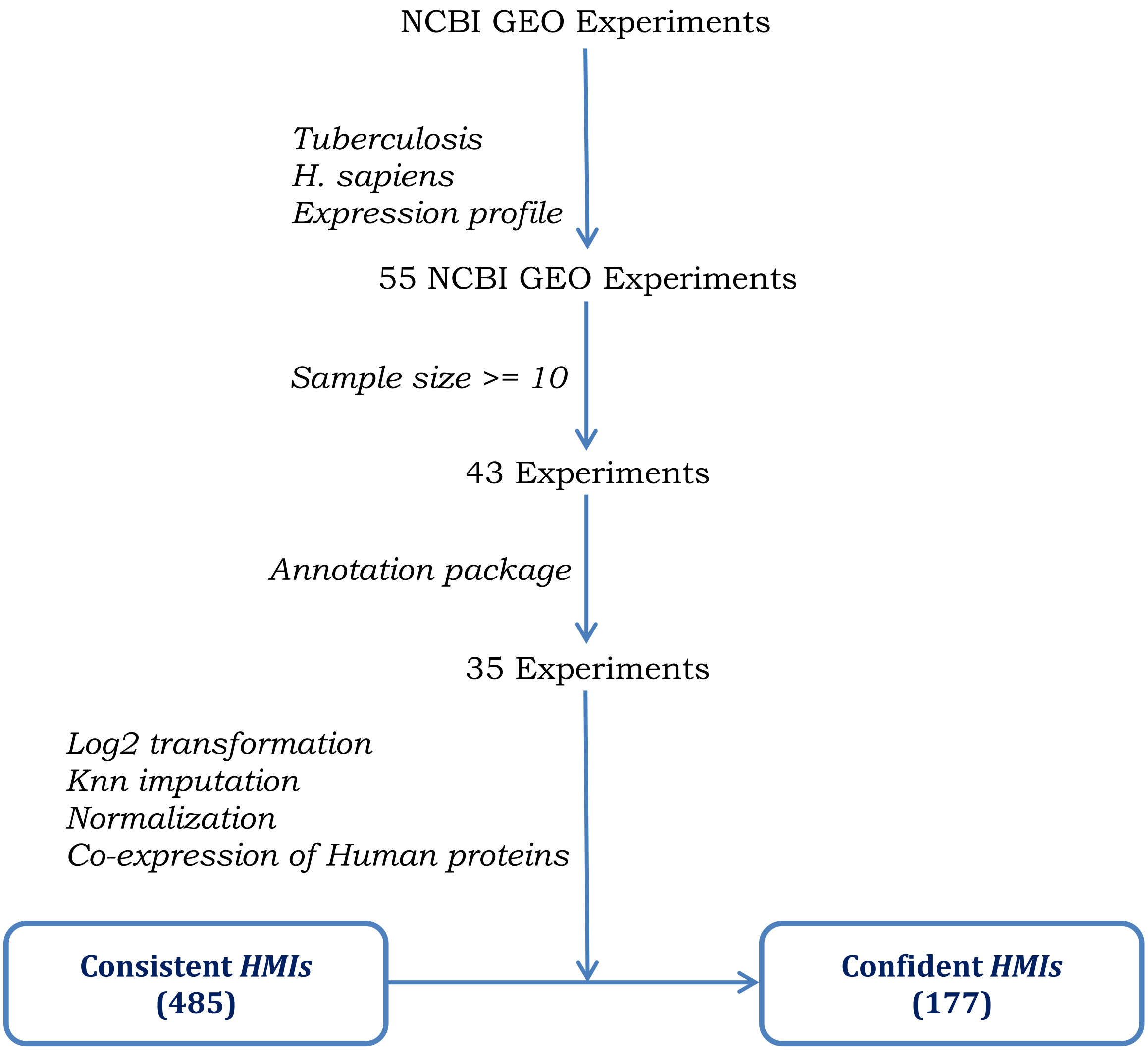


Figure 4.6 Procedure for co-expression analysis of *Human genes*

The number of confident *HMIs* detected through functional annotation is 177. The number of confident *HMIs* detected through co-expression analysis is 316. Of the 485 consistent *HMIs*, a total of 419 *HMIs* are detected as confident interactions.

4.3.5 Database of annotated microarray experiments (DAME)

Database of annotated microarray experiments (DAME) is a database developed in-house during the co-expression analysis of the *Human* proteins in the *HMIs*. It is well known that large volumes of microarray data are available at NCBI GEO. The microarray data is generally analyzed and annotated manually which is a tedious process. Currently, there is no web resource for retrieval of such data. Therefore, in the present study, a web resource is developed to store the analyzed and annotated information of the microarray data. This is achieved with the development of web pages using PHP, HTML, JAVA scripts and Apache 2.0 webserver. rApache [34] is used to support the webserver with the R [35] environment. Perl module is integrated for easy operations and R module is enabled to run R commands on the webserver. Further, a MySQL database is integrated for the retrieval of experimental information related to the microarray data.

Currently, DAME stores the analyzed and annotated information of the experiments related to TB, AIDS, dengue, malaria, and diabetes. Differentially expressed genes, cluster of samples, classification maps and co-express networks are produced from the microarray experimental data using R scripts. Gene symbols, functional annotation [27] and pathways [36] are generated using R scripts along with the Bioconductor packages. A database is created that stores all the information related to the microarray experiments. All the data generated is stored in DAME. A user can retrieve the experimental information of the microarray data from Experiments menu, analyzed information from Analyze menu and annotated information from the Annotate menu. The following **Figure 4.7** is a flow chart for the application of DAME database.

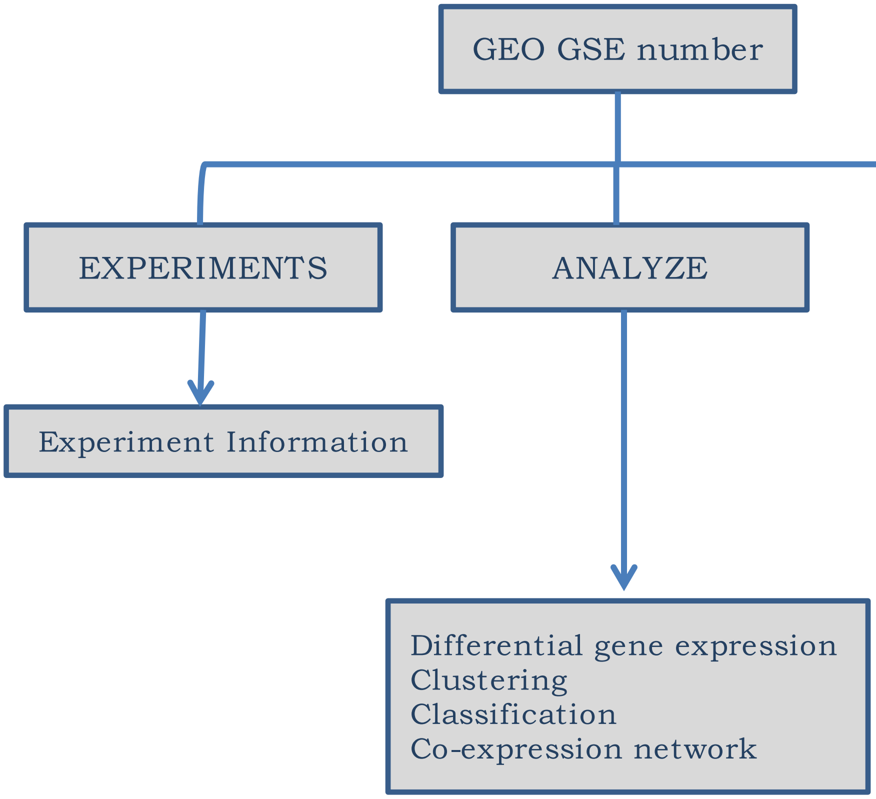


Figure 4.7 Flow-chart for the application of DAME

In addition to the retrieval of stored data, the user can update the set of genes of interest by changing the parameters for the differentially expressed genes or by directly searching the genes of interest in the database. The updated genes set can subsequently be used to customize the classification map, co-expression network, functional annotations and pathways through the database web interface. Therefore, the information obtained from the database can aid the user in establishing gene functions [27, 37] and pathways [36, 38-39]. It can support disease subtype classifications [39-40] and progression studies. It can also motivate the user to investigate disease specific molecular markers and the new gene targets.

4.3.5.1 An example for DAME application

To demonstrate the application of DAME, experiment series number GSE42827 [41] obtained from GEO is considered as an example. The experiment is used to compare the whole blood transcriptional signatures of pneumonia patients before and after antibiotic treatment. The information such as title, samples, platforms and contributors for this experiment are retrieved from DAME database by entering the GSE number in the search text box of EXPERIMENTS menu. The experiment is found to have 10 samples with 5 samples of before antibiotic treatment and with other 5 samples of after antibiotic treatment. Clustering of samples is readily visualized from the cluster plot obtained from ANALYZE → Clustering. The following **Figure 4.8 (s)** is a cluster plot generated for the present example. From the figure, it is clear that GSM1050877, GSM1050878, GSM1050879, GSM1050880 and GSM1050885 are the controls samples with before antibiotic treatment while GSM1050876, GSM1050881, GSM1050882, GSM1050883 and GSM1050884 are the test samples with after antibiotic treatment.

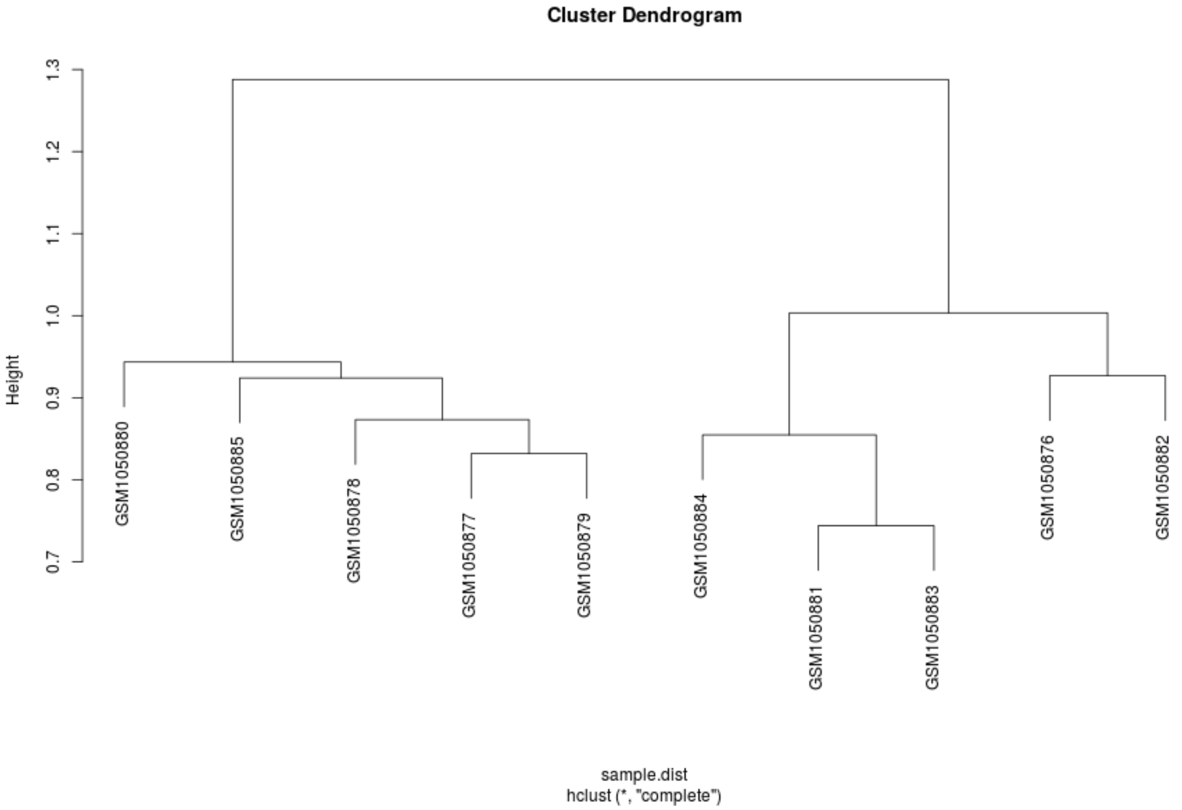


Figure 4.8 (a) Clustering of samples

Top 10 differentially expressed genes in the experiment are obtained from ANALYZE → DGE. A user can update the genes of interest from the differentially expressed genes list using the Update button or directly search the genes of interest in the database. Heatmap of the differentially expressed genes is obtained from ANALYZE → Classification. The following **Figure 4.8 (b)** is the heatmap of classification of the differentially expressed genes. The green color in the heatmap indicates the down-regulation of the genes while red color indicates the up-regulation of the genes. The user can also update the heatmap for the genes of interest.

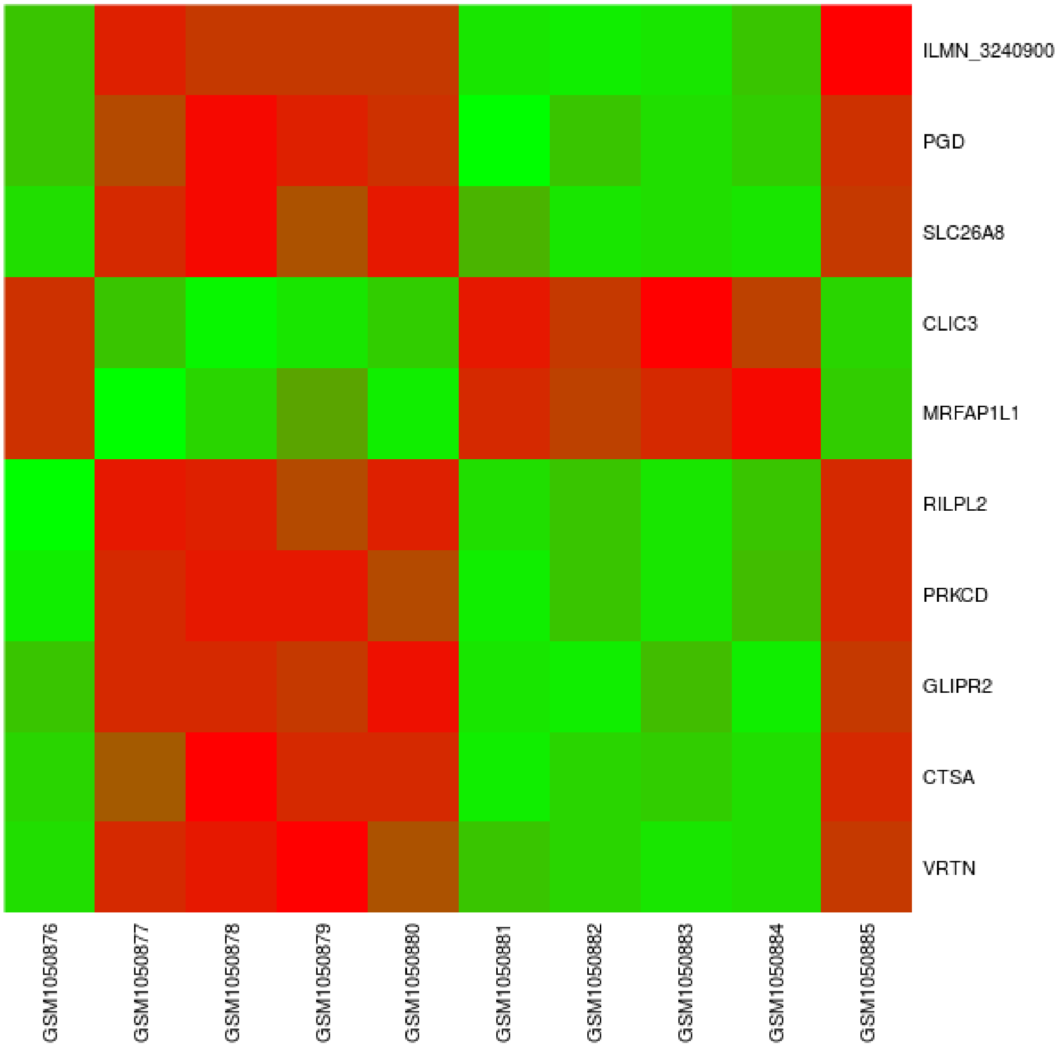


Figure 4.8 (b) Classification of differentially expressed genes

Co-expression network of the differentially expressed genes is obtained from ANALYZE → Coexp_ntwk. The following **Figure 4.8 (c)** is the network plot of differentially expressed genes. When the correlation of expression profiles of two differentially expressed genes is minimum 0.8, a link is formed between the two genes in the network. The user can also update the network for the genes of interest.

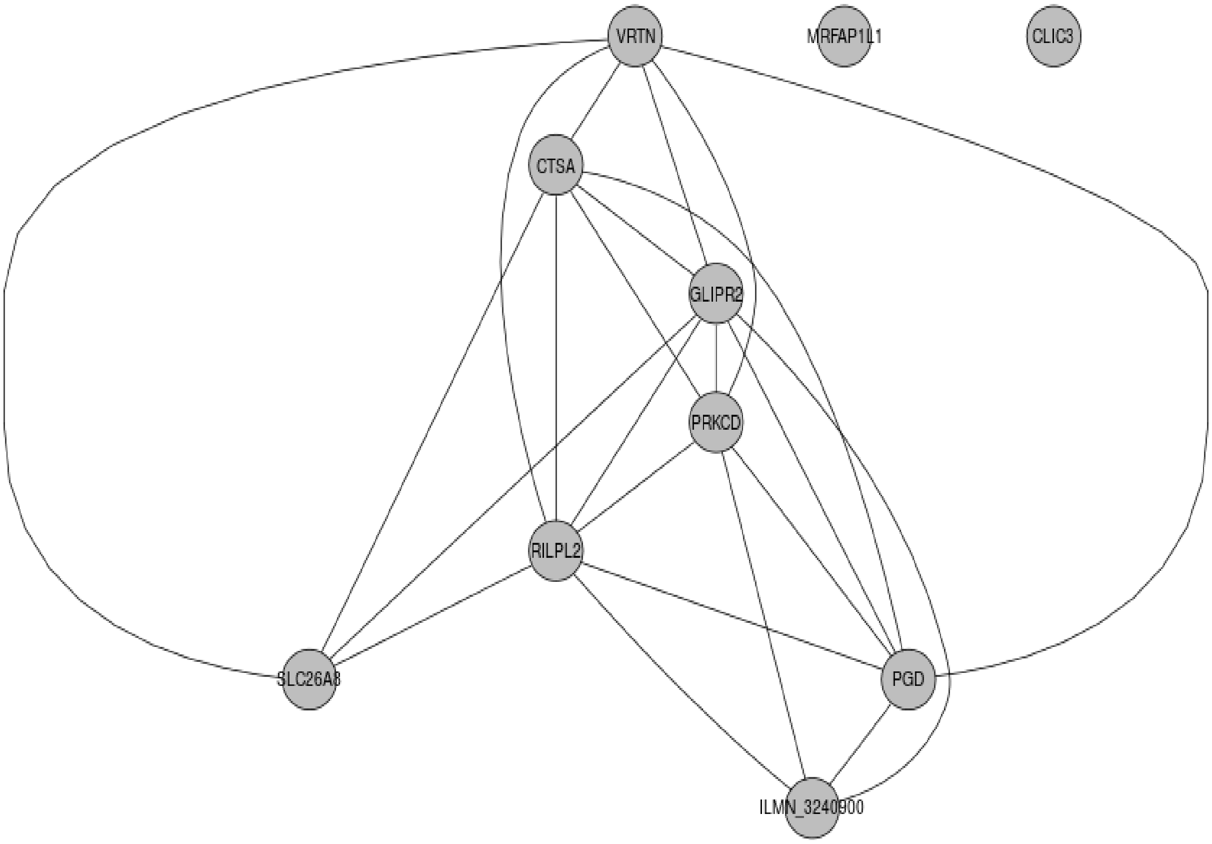
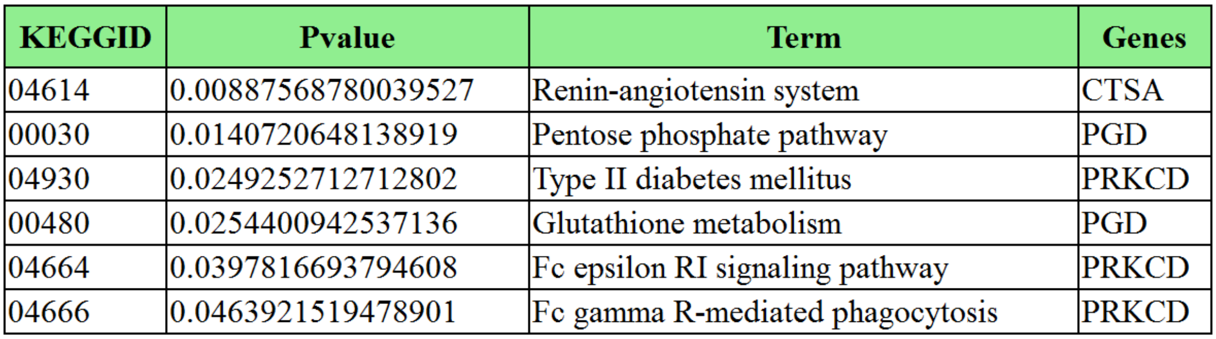


Figure 4.8 (c) Co-expression networks of differentially expressed genes

Gene symbols of the probes in the microarray experiment is obtained from ANNOTATE → Gene Symbols. Annotations of the differentially expressed genes are obtained from ANNOTATE → Annotate. The user can obtain the annotated information for the genes of interest. Ontology terms related to biological processes, molecular functions and cellular components of the differentially expressed genes are obtained from ANNOTATE → GO Terms. Pathways of the differentially expressed genes are obtained from ANNOTATE → KEGG

Pathways. The user can update the ontology terms and the pathways for the genes of interest. Both the ontology terms and the pathways are viewed as HTML reports and the table format. In the table format, genes of the ontology terms and the pathways are also revealed. The following **Table 4.3** represents the KEGG pathways for the differentially expressed genes.

Table 4.3 Table of KEGG pathways



KEGGID	Pvalue	Term	Genes
04614	0.00887568780039527	Renin-angiotensin system	CTSA
00030	0.0140720648138919	Pentose phosphate pathway	PGD
04930	0.0249252712712802	Type II diabetes mellitus	PRKCD
00480	0.0254400942537136	Glutathione metabolism	PGD
04664	0.0397816693794608	Fc epsilon RI signaling pathway	PRKCD
04666	0.0463921519478901	Fc gamma R-mediated phagocytosis	PRKCD

4.4 Conclusions

For the first time, I2H is employed in the present study to predict host-pathogen protein interactions. A set of physically interacting *HMIs* is predicted using I2H. Another set of *HMIs* is predicted using Interlog method. *HMIs* that are consistently predicted by both the methods are obtained. The consistently predicted *HMIs* are filtered using functional annotations and co-expression analysis to generate confident *HMIs*. With this approach, a total of 419 confident *HMIs* are obtained. These *HMIs* could prove essential in elucidating the dynamics of Human-*M.TB* protein interactions as well as the infection and persistence mechanisms of the pathogen. DAME, a database of annotated microarray experiments is developed to store all the analyzed and annotated information of the microarray experiments related to TB, HIV and other diseases. Information from the database can influence the current TB research for investigating disease novel molecular markers and therapeutic studies.

4.5 References

1. Smith I, "*Mycobacterium tuberculosis* pathogenesis and molecular determinants of virulence", Clin Microbiol Rev 2003, 16(3): 463-496.
2. Mitchell G, Chen C, Portnoy DA, "Strategies used by bacteria to grow in macrophages", Microbiol Spectr 2016, 4(3).
3. Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N, Klitgord N, Simon C, Boxem M, Milstein S, Rosenberg J, Goldberg DS, Zhang LV, Wong SL, Franklin G, Li S, Albala JS, Lim J, Fraughton C, Llamosas E, Cevik S, Bex C, Lamesch P, Sikorski RS, Vandenhaute J, Zoghbi HY, Smolyar A, Bosak S, Sequerra R, Doucette-Stamm L, Cusick ME, Hill DE, Roth FP, Vidal M, "Towards a proteome-scale map of the human protein-protein interactions network", Nature 2005, 437(7062): 1173–1178.
4. Li S, Armstrong CM, Bertin N, Ge H, Milstein S, Boxem M, Vidalain PO, Han JD, Chesneau A, Hao T, Goldberg DS, Li N, Martinez M, Rual JF, Lamesch P, Xu L, Tewari M, Wong SL, Zhang LV, Berriz GF, Jacotot L, Vaglio P, Reboul J, Hirozane-Kishikawa T, Li Q, Gabel HW, Elewa A, Baumgartner B, Rose DJ, Yu H, Bosak S, Sequerra R, Fraser A, Mango SE, Saxton WM, Strome S, Van Den Heuvel S, Piano F, Vandenhaute J, Sardet C, Gerstein M, Doucette-Stamm L, Gunsalus KC, Harper JW, Cusick ME, Roth FP, Hill DE, Vidal M, "A map of the interactome network of the metazoan *C. elegans*", Science 2004, 303(5657): 540–543.
5. Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao YL, Ooi CE, Godwin B, Vitols E, Vijayadamodar G, Pochart P, Machineni H, Welsh M, Kong Y, Zerhusen B, Malcolm R, Varrone Z, Collis A, Minto M, Burgess S, McDaniel L, Stimpson E, Spriggs F, Williams J, Neurath K, Ioime N, Agee M, Voss E, Furtak K, Renzulli R, Aanensen N, Carrolla S, Bickelhaupt E, Lazovatsky Y, DaSilva A, Zhong J, Stanyon CA, Finley RL Jr, White KP, Braverman M, Jarvie T, Gold S, Leach M, Knight J, Shimkets RA, McKenna

MP, Chant J, Rothberg JM, “A protein interaction map of *Drosophila melanogaster*”, Science 2003, 302(5651): 1727–1736.

6. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K, Yang L, Wolting C, Donaldson I, Schandorff S, Shewnarane J, Vo M, Taggart J, Goudreault M, Muskat B, Alfarano C, Dewar D, Lin Z, Michalickova K, Willems AR, Sassi H, Nielsen PA, Rasmussen KJ, Andersen JR, Johansen LE, Hansen LH, Jespersen H, Podtelejnikov A, Nielsen E, Crawford J, Poulsen V, Sørensen BD, Matthiesen J, Hendrickson RC, Gleeson F, Pawson T, Moran MF, Durocher D, Mann M, Hogue CW, Figeys D, Tyers M, “Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry”, Nature 2002, 415(6868): 180–183.
7. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y, “A comprehensive two-hybrid analysis to explore the yeast protein interactome”, Proc Natl Acad Sci U S A 2001, 98(8): 4569–4574.
8. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, Qureshi-Emili A, Li Y, Godwin B, Conover D, Kalbfleisch T, Vijayadamodar G, Yang M, Johnston M, Fields S, Rothberg JM, “A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae”,* Nature 2000, 403(6770): 623–627.
9. Venkatesan K, Rual JF, Vazquez A, Stelzl U, Lemmens I, Hirozane-Kishikawa T, Hao T, Zenkner M, Xin X, Goh KI, Yildirim MA, Simonis N, Heinzmann K, Gebreab F, Sahalie JM, Cevik S, Simon C, de Smet AS, Dann E, Smolyar A, Vinayagam A, Yu H, Szeto D, Borick H, Dricot A, Klitgord N, Murray RR, Lin C, Lalowski M, Timm J, Rau K, Boone C, Braun P, Cusick ME, Roth FP, Hill DE, Tavernier J, Wanker EE, Barabasi AL, Vidal M, “An empirical framework for binary interactome mapping”, Nat methods 2008, 6(1): 83-90.
10. Fabregat A, Sidiropoulos K, Garapati P, Gillespie M, Hausmann K, Haw R, Jassal B, Jupe S, Korninger F, McKay S, Matthews L, May B, Milacic M, Rothfels K, Shamovsky V, Webber M, Weiser J, Williams M, Wu G, Stein L,

Hermjakob H, D'Eustachio P, "The Reactome pathway Knowledgebase", Nucleic Acids Res 2016; 44(D1): D481-D487.

11. Dandekar T, Snel B, Huynen M, Bork P, "Conservation of gene order: a fingerprint of proteins that physically interact", Trends Biochem Sci 1998, 23(9): 324-328.
12. Korbel JO, Jensen LJ, von Mering C, Bork P, "Analysis of genomic context: prediction of functional associations from conserved bidirectionally transcribed gene pairs", Nat Biotechnol 2004, 22(7): 911-917.
13. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO, "Assigning protein functions by comparative genome analysis: protein phylogenetic profiles", Proc Natl Acad Sci U S A 1999, 96(8): 4285-4288.
14. Walhout AJ, Sordella R, Lu X, Hartley JL, Temple GF, Brasch MA, Thierry-Mieg N, Vidal M, "Protein interaction mapping in *C. elegans* using proteins involved in vulval development", Science 2000, 287(5450): 116-122.
15. Rapanoel HA, Mazandu GK, Mulder NJ, "Predicting and analyzing interactions between *Mycobacterium tuberculosis* and its human host", PLoS One 2013, 8(7): e67472.
16. Pazos F, Valencia A, "*In silico* two-hybrid system for the selection of physically interacting protein pairs", Proteins 2002, 47(2): 219-227.
17. Gobel U, Sander C, Schneider R, Valencia A, "Correlated mutations and residue contacts in proteins", Proteins 1994, 18(4): 309-317.
18. Jothi R, Cherukuri PF, Tasneem A, Przytycka TM, "Co-evolutionary analysis of domains in interacting proteins reveals insights into domain-domain interactions mediating protein–protein interactions", J Mol Biol 2006, 362(4): 861–875.
19. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ, "Basic Local alignment search tool", J Mol Biol 1990, 215: 403-410.
20. Fulton DL, Li YY, Laird MR, Horsman BG, Roche FM, Brinkman FS, "Improving the specificity of high-throughput orthologs prediction", BMC Bioinformatics 2006, 7: 270.

21. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Soding J, Thompson JD, Higgins DG, “Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega”, Mol Syst Biol 2011, 7: 539.
22. McLachlan AD, “Test for comparing related amino acid sequences”, J Mol Biol 1971, 61: 409–424.
23. Xenarios I, Salwínski L, Duan XJ, Higney P, Kim SM, Eisenberg D, “DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions”, Nucleic Acids Res 2002, 30(1): 303-305.
24. Lopez Y, Nakai K, Patil A, “HitPredict version 4: comprehensive reliability scoring of physical protein–protein interactions from more than 100 species”, Database (Oxford) 2015, 2015: bav117.
25. 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, 33: D433-D437.
26. Matthews LR, Vaglio P, Reboul J, Ge H, Davis BP, Garrels J, Vincent S, Vidal M, “Identification of potential interactions networks using sequence-based searches for conserved protein-protein interactions or interlogs”, Genome Res 2001, 11(12): 2120-2126.
27. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G, “Gene ontology: tool for the unification of biology. The Gene Ontology Consortium”, Nat Genet 2000, 25(1): 25-29.
28. Mi H, Poudel S, Muruganujan A, Casagrande JT, Thomas PD, “PANTHER version 10: expanded protein families and functions, and analysis tools”, Nucleic Acids Res 2016, 44(D1): D336-D342.
29. Dyer MD, Murali TM, Sobral BW, “Computational prediction of host-pathogen protein-protein interactions”, Bioinformatics 2007, 23(13): i159-i166.

30. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A, Lee H, Zhang N, Robertson CL, Serova N, Davis S, Soboleva A, "NCBI GEO: archive for functional genomics data sets—update", Nucleic Acids Res 2013, 41(Database issue): D991-D995.
31. Wang Z, Arat S, Magid-Slav M, Brown JR, "Meta-analysis of human gene expression in response to *Mycobacterium tuberculosis* infection reveals potential therapeutic targets", BMC Syst Biol 2018, 12(1): 3.
32. Lee HK, Hsu AK, Sajdak J, Qin J, Pavlidis P, "Coexpression analysis of human genes across many microarray data sets", Genome Res 2004, 14(6): 1085-1094.
33. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, Zhang J, "Bioconductor: open software development for computational biology and bioinformatics", Genome Biol 2004, 5(10): R80.
34. Horner J, "rApache: Web application development with R and Apache", 2013, url = http://www.rapache.net/.
35. R Core Team, "R: A Language and Environment for Statistical Computing", R Foundation for Statistical Computing, Vienna, Austria 2013.
36. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K, "KEGG: new perspectives on genomes, pathways, diseases and drugs", Nucleic Acids Res 2017, 45(D1): D353-D361.
37. van Dam S, Võsa U, van der Graaf A, Franke L, de Magalhaes JP, "Gene co-expression analysis for functional classification and gene-disease predictions", Brief Bioinform 2018, 19(4): 575-592.
38. Tomfohr J, Lu J, Kepler TB, "Pathway level analysis of gene expression using singular value decomposition", BMC Bioinformatics 2005, 6: 225.

39. Wanggou S, Feng C, Xie Y, Ye L, Wang F, Li X, "Sample Level Enrichment Analysis of KEGG Pathways Identifies Clinically Relevant Subtypes of Glioblastoma", J Cancer 2016, 7(12): 1701-1710.

40. Rapin N, Bagger FO, Jendholm J, Mora-Jensen H, Krogh A, Kohlmann A, Thiede C, Borregaard N, Bullinger L, Winther O, Theilgaard-Monch K, Porse BT, "Comparing cancer vs normal gene expression profiles identifies new disease entities and common transcriptional programs in AML patients", Blood 2014, 123(6): 894-904.

41. Bloom CI, Graham CM, Berry MP, Rozakeas F, Redford PS, Wang Y, Xu Z, Wilkinson KA, Wilkinson RJ, Kendrick Y, Devouassoux G, Ferry T, Miyara M, Bouvry D, Valeyre D, Gorochov G, Blankenship D, Saadatian M, Vanhems P, Beynon H, Vancheeswaran R, Wickremasinghe M, Chaussabel D, Banchereau J, Pascual V, Ho LP, Lipman M, O'Garra A, "Transcriptional blood signatures distinguish pulmonary tuberculosis, pulmonary sarcoidosis, pneumonias and lung cancers", PLoS One 2013, 8(8): e70630.

CHAPTER 5

Construction and Analysis of *Human* and *M. tuberculosis* Protein Interactions Network

5.1 Introduction

MTB is an ancient and devastating pathogen dwelling in the *Human* macrophages. It persists in the macrophages [1-2] interacting and influencing several of the *Human* proteins [3-5]. A network of such relations could reveal several mysteries related to infection and survival of the pathogen. Since the efficient experimental methods for sketching of host pathogen protein interactions are unavailable [6], computational methods are employed [7]. Recently, implementation of the computation methods has disclosed few surface proteins of *MTB* that interacts with intracellular proteins of the *Human* [8]. Further, some proteins are identified that trigger *Human* immune response and interact with HIV [9]. A *HMIs* network of such relations is very small with a limited number of proteins. Therefore, the *HMI* network generated from the whole genomic context is essential in explicating the pathogen mechanisms during infection and co-infection with other pathogens. The pathogen makes direct contact with the surface proteins of *Human* cells during infection [10]. Since the sequence based methods are more universal and reliable [11-12], physical interactions predicted at the genome level are vital for the development of a *HMI*s network. Such a network not only helps in understanding the virulence of the pathogen but also in identifying suitable drug targets [13].

In the present study, a network is constructed from the confident *HMIs* generated from the previous Chapter 4. The *HMIs* network is analyzed to deduce a set of highly interacting *MTB* proteins and the *Human* proteins. Interactions of a hypothetical protein of *MTB* with the proteins of *Human* are identified. *MTB* proteins that are critical during infection and progression of the disease are detected. Further, a set of *MTB* proteins which could be potential drug targets is uncovered.

5.2 Materials and Methods

5.2.1 *Human* and *M. tuberculosis* protein interactions network

Confident *HMIs* generated from the previous Chapter 4 are used to construct a *HMIs* network. It is accomplished with the implementation of Cytoscape software [14].

Topological properties such as the number of *Human* and *MTB* nodes (proteins), total number of edges (interactions), highly connected *MTB* nodes and *Human* nodes are deduced from the interactions network. Further, hypothetical proteins and their interactions in the interactions network are identified to understand their critical role during TB infection.

5.2.2 Pathways

Human proteins are separated from the *MTB* proteins in the *HMIs* network. Gene Ids of these proteins are retrieved from *Human* proteins table obtained through NCBI genome server. The gene ids along with pink background and blue foreground color coding are loaded on KEGG server [15] to generate a set of *Human* proteins that are known to play key roles in TB pathways. An interaction sub-network of these *Human* proteins is derived from the *HMIs network* and used to understand the pathways in the course of infection and progression of the disease. Further, *MTB* nodes in the sub-network are inspected for the potential therapeutic properties from the existing literature.

5.3 Results and Discussion

5.3.1 *Human-M. tuberculosis* protein interactions network

Confident *HMIs* are used to build a *HMIs* network. In the *HMIs* network, the *Human* proteins are represented by green colored circular balls or the *Human* nodes and *MTB* proteins are represented by pink colored circular balls or the *MTB* nodes. Interactions between the *Human* nodes and the *MTB* nodes are represented by black colored lines or edges. The following **Figure 5.1** represents the network of *HMIs.* The number of *Human* nodes in the *HMIs* network is found to be 251 while the number of *MTB* nodes is observed as 20. The *MTB* nodes or proteins are examined in the existing literature and databases. It is observed that 11 of them are previously revealed to play crucial roles during infection [8-9]. Therefore, from the present study, 9 *MTB* proteins are reported as the new and significant proteins to involve in infection and pathogenesis of TB. The following **Table 5.1 (a) and Table 5.1 (b)** list the known and new *MTB* proteins in the *HMIs*.

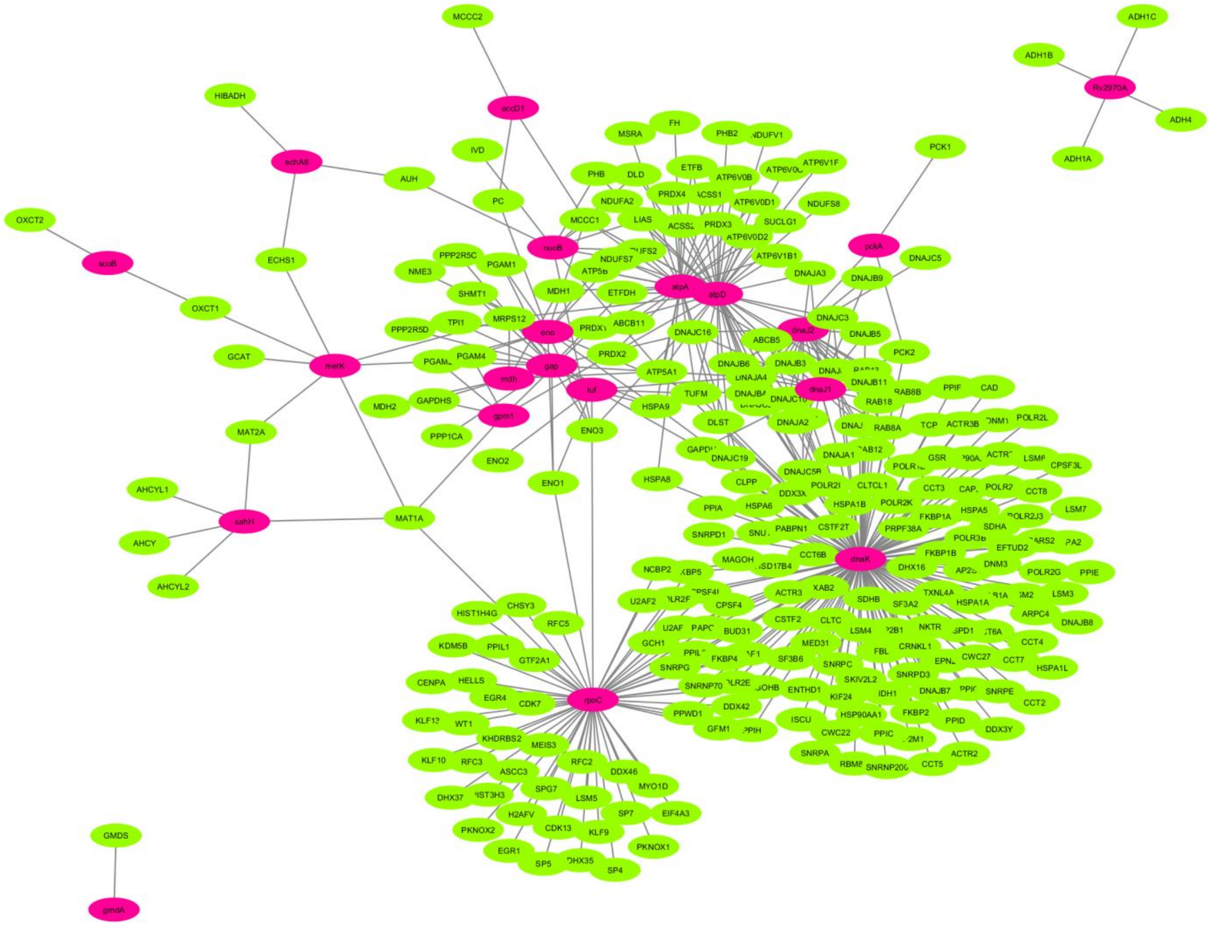


Figure 5.1 The *HMIs* network

Table 5.1 (a) Known *MTB* proteins in *HMIs*

1.	atpA	7.	gpm1
2.	atpD	8.	metK
3.	dnaJ1	9.	rpoC
4.	dnaK	10.	scoB
5.	Eno	11.	tuf
6.	gap		

Table 5.1 (b) New *MTB* proteins in *HMIs*

1.	accD1	6.	nuo
2.	dnaJ2	7.	pckA
3.	echA8	8.	Rv2970A
4.	gmdA	9.	sahH
5.	mdh		

In the interactions network, the top highest degree [16] nodes of *MTB* are observed to be dnaK with a value of 147, rpoC with 59, atpD with 47, atpA with 29 and dnaJ1 with a value of 25. Number of interactions for these nodes is measured to be more than 70 percent of the total number of interactions in the *HMIs* network. Therefore, the removal of such nodes causes fatality [16]. Top highest degree nodes of *Human* are observed to be DNJA4 with a value of 7, ENO, TUFM and MRPS12 with a value of 6 and ATP5A1 and HSPA9 with a value of 5. These nodes are observed to contribute more than 8 percent of the total number of interactions in the *HMIs* network. Therefore, the study of these nodes could help in understanding the regulatory mechanisms involved during TB disease.

5.3.2 Hypothetical proteins

In the interactions network, Rv2970A, a conserved hypothetical protein of *MTB* is found to be connected with 4 *Human* proteins, viz., ADH1B, ADH1C, ADH1A and ADH4. These *Human* proteins are different isoforms of alcohol dehydrogenase enzyme. Recently, it is demonstrated that alcohol metabolizing enzymes are genetically associated with the risk of causing TB *[17]*. Therefore, the interactions of Rv2970A with the different isoforms of alcohol dehydrogenase enzyme indicate that it has properties similar to the enzyme and play an important role during TB infection. The following **Figure 5.2** is as sub-network of Rv2970A in the *HMIs* network.

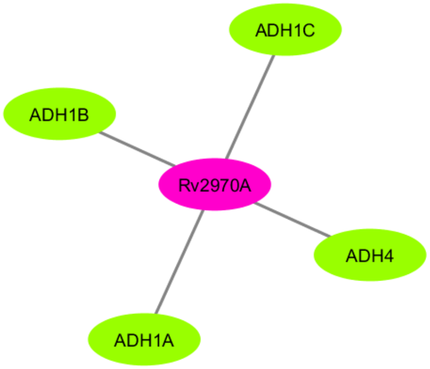


Figure 5.2 The sub-network of Rv2970A

5.3.3 Pathway analysis

Gene Ids of 251 *Human* proteins present in the *HMIs* network are extracted from the protein table and used to map in KEGG database *[15]*. The following **Figure 5.3** is the TB pathway obtained from KEGG database.

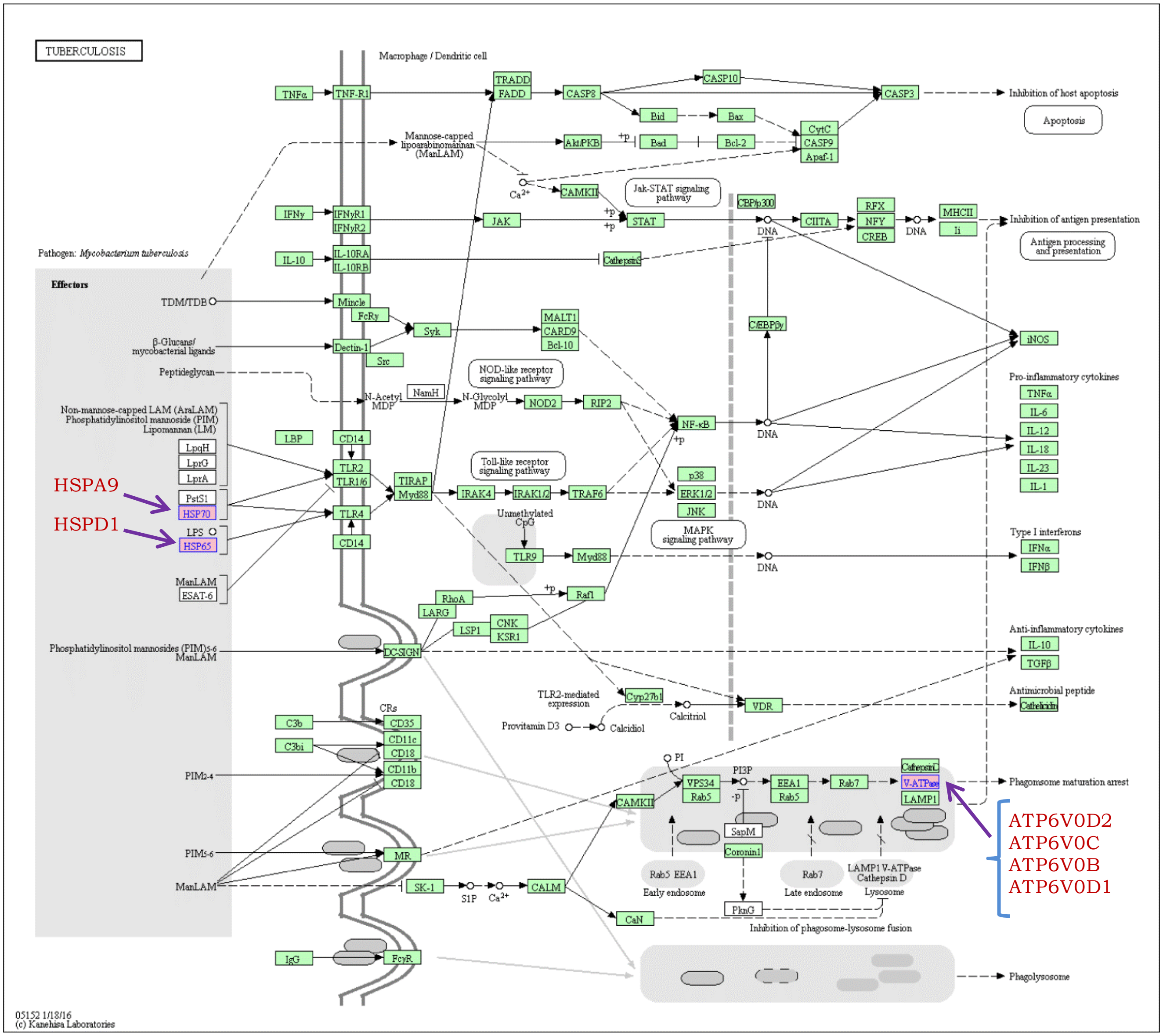


Figure 5.3 The KEGG pathway of tuberculosis

It is observed from **Figure 5.3** that six *Human* proteins viz., HSPA9, HSPD1, ATP6VOB, ATP6VOC, ATP6VOD1 and ATP6VOD2 of the *HMIs* network are involved in TB pathways. The figure also reveals the mechanism of interactions in the TB pathway. It is perceived that HSPA9 which encodes a member of HSP70 family protein [18], signals through TLR2 and TLR4 (toll like receptor proteins) [19-20] and HSPD1 which encodes a member of HSP60 family protein

[21] signals through TLR4 [19]. These TLR proteins are present on macrophage or dendritic cell membrane. It is reported that HSPA9 plays role in proliferation and maintenance of the cells [22] while HSPD1 plays role in survival and inflammatory responses [23]. Further, ATP6V0C, ATP6V0B, ATP6V0D1 and ATP6V0D2 are the V-type proton ATPase subunit isoforms [24] that involve in phagosome maturation arrest [25-27] which cause the inhibition of antigen presentation. Thus, these *Human* proteins of *HMIs* network provide insights into the TB infection and successive pathways.

5.3.4 TB sub-network

A sub-network is extracted from the *HMIs* network for the interactions of the 6 *Human* proteins involved in the TB pathways. The following **Figure 5.4** represents the sub-network of TB.

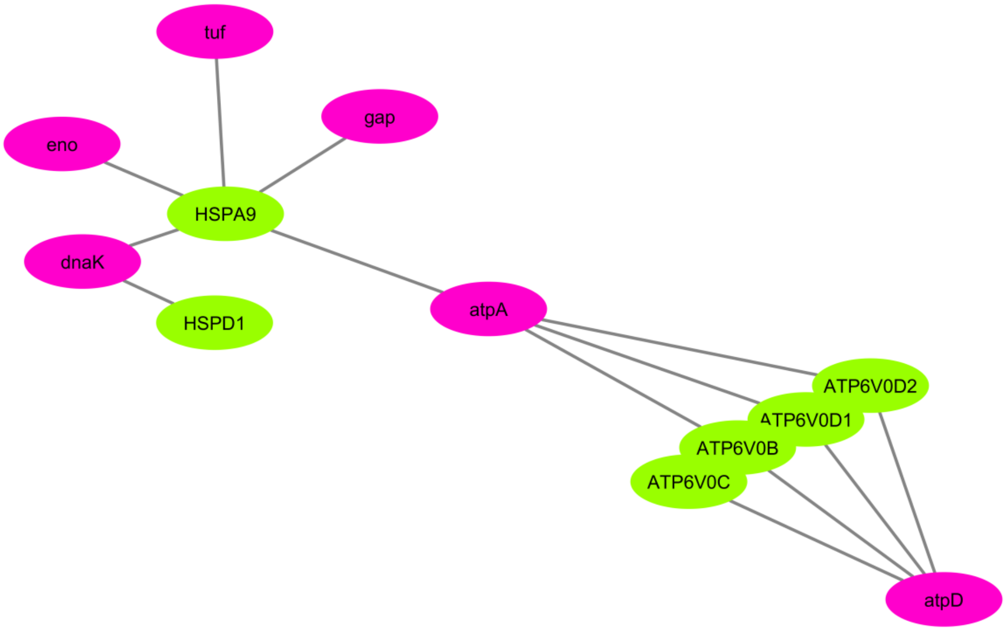


Figure 5.4 The TB sub-network from the *HMIs*

It is indicated from the figure that dnaK chaperone protein of *MTB* interacts with HSPD1 and HSPA9 of the *Human* proteins to trigger TLR signaling pathways of TB. It is found that eno, tuf and gap proteins of *MTB* also trigger TLR signaling pathways through the interaction with HSPA9. It is observed that atpD protein of *MTB* interact with different isoforms of V-ATPase protein subunits and initiate the process of phagosome maturation arrest. Further, it is detected that atpA protein of *MTB* plays role in both initiating the TLR signaling pathways and processing the phagosome maturation arrest. Therefore, these *MTB* proteins can be useful in understanding the pathogen mechanisms as well as in developing new therapeutic intervention strategies.

5.3.5 Significance of TB sub-network

MTB proteins in the TB sub-network are compared with the DEG database [28]. It is observed that all of the *MTB* proteins in the TB sub-network are the essential gene products. The *MTB* proteins are also compared with the Tuberculist database [29]. It is observed that dnaK play role in virulence, detoxification and adaptation processes, eno, gap, atpA and atpD are involved in intermediary metabolism and respiration processes, and tuf is implicated in information pathways. The following **Table 5.2** represents the significance of the *MTB* proteins in the TB sub-network. Further, atpD, atpA, dnaK and eno are the known drug targets [8]. Therefore, in the present study, tuf and gap are reported as the new and potential *MTB* proteins which could be explored further for their ability to operate as potential drug targets.

Table 2. Significance of *MTB* proteins in the TB sub-network

Sr. No.	Protein	Degree	Essentiality	Tuberculist Pathways
1	atpD	High	Essential	Intermediary Metabolism and Respiration
2	atpA	-	Essential	Intermediary Metabolism and Respiration
3	dnaK	High	Essential	Virulence, Detoxification and Adaptation
4	eno	-	Essential	Intermediary Metabolism and Respiration
5	tuf	-	Essential	Information Pathways
6	gap	-	Essential	Intermediary Metabolism and Respiration

5.4 Conclusions

The protein interactions network of *HMIs* is produced from the confident *HMIs*. The *HMIs* network is found to contain 419 interactions among 20 *MTB* proteins and 251 *Human* proteins. The highest degree *MTB* protein in the *HMIs* network is observed to be dnaK followed by rpoC and atpD proteins. The highest degree *Human* protein is observed to be DNAJ4. Rv2970A is a hypothetical protein found to interact with different isoforms of *Human* alcohol dehydrogenase enzyme. It is detected that six *Human* proteins of *HMIs* network are involved in TB pathways. A sub-network of these proteins has predicted that six *MTB* proteins play a crucial role during initiation and progression of TB disease. Further, these *MTB* proteins are observed to be essential gene products and involve in various functional categories. Four of these proteins are already known to be drug targets. Therefore, two proteins, "tuf" and "gap" can be explored for their proficiency as drug targets in the TB research.

5.5 References

1. Kumar D, Nath L, Kamal MA, Varshney A, Jain A, Singh S, Rao KV, "Genome-wide analysis of the host intracellular network that regulates survival of *Mycobacterium tuberculosis*", Cell 2010, 140(5): 731-743.
2. Jayaswal S, Kamal MA, Dua R, Gupta S, Majumdar T, Das G, Kumar D, Rao KV, "Identification of host-dependent survival factors for intracellular *Mycobacterium tuberculosis* through an siRNA screen", PLoS Pathog 2010, 6(4): e1000839.
3. Basu SK, Kumar D, Singh DK, Ganguly N, Siddiqui Z, Rao KV, Sharma P, "*Mycobacterium tuberculosis* secreted antigen (MTSA-10) modulates macrophage function by redox regulation of phosphatases", FEBS J 2006, 273(24): 5517-5534.
4. Raghavan S, Manzanillo P, Chan K, Dovey C, Cox JS, "Secreted transcription factor controls *Mycobacterium tuberculosis* virulence", Nature 2008, 454(7205): 717-721.

5. Mitchell G, Chen C, Portnoy DA, "Strategies Used by Bacteria to Grow in Macrophages", Microbiol Spectr 2016, 4(3).
6. Venkatesan K, Rual JF, Vazquez A, Stelzl U, Lemmens I, Hirozane-Kishikawa T, Hao T, Zenkner M, Xin X, Goh KI, Yildirim MA, Simonis N, Heinzmann K, Gebreab F, Sahalie JM, Cevik S, Simon C, de Smet AS, Dann E, Smolyar A, Vinayagam A, Yu H, Szeto D, Borick H, Dricot A, Klitgord N, Murray RR, Lin C, Lalowski M, Timm J, Rau K, Boone C, Braun P, Cusick ME, Roth FP, Hill DE, Tavernier J, Wanker EE, Barabási AL, Vidal M, "An empirical framework for binary interactome mapping", Nat Methods 2009, 6(1): 83-90.
7. Dyer MD, Murali T, Sobral BW, "Computational prediction of host-pathogen protein–protein interactions", Bioinformatics 2007, 23: i159-i166.
8. Rapanoel HA, Mazandu GK and Mulder NJ, "Predicting and analyzing interactions between *Mycobacterium tuberculosis* and its human host", PLOS One 2013, 8: e67472.
9. Cui T, Li W, Liu L, Huang Q, He ZG, "Uncovering New Pathogen-Host Protein-Protein Interactions by Pairwise Structure Similarity", PLoS One 2016, 11(1): e0147612.
10. Smith I, "*Mycobacterium tuberculosis* pathogenesis and molecular determinants of virulence", Clin Microbiol Rev 2003, 16: 463-496.
11. Shen J, Zhang J, Luo X, Zhu W, Yu K, Chen K, Li Y, Jiang H, "Predicting protein–protein interactions based only on sequences information", PNAS 2007, 104(11): 4337-4341.
12. Bharne D, Naresh D, Vindal V, "Inferring protein interactions network of *Mycobacterium tuberculosis H37Rv* using sequence information", Res J Life Sci Bioinform Pharm Chem Sci 2018, 4(6): 57-64.
13. Hase T, Tanaka H, Suzuki Y, Nakagawa S, Kitano H, "Structure of protein interactions networks and their implications on drug design", PLoS Comput Biol 2009, 5(10): e1000550.

14. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T, “Cytoscape: a software environment for integrated models of biomolecular interactions networks”, Genome Res 2003, 13(11): 2498-2504.
15. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K, “KEGG: new perspectives on genomes, pathways, diseases and drugs”, Nucleic Acids Res 2017, 45(D1): D353-D361.
16. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL, “Lethality and centrality in protein networks”, Nature 2001, 411: 41-42.
17. Park SK, Park CS, Lee HS, Park KS, Park BL, Cheong HS, Shin HD, “Functional polymorphism in aldehyde dehydrogenase-2 gene associated with risk of TB”, BMC Med Genet 2014, 15: 40.
18. Domanico SZ, DeNagel DC, Dahlseid JN, Green JM, Pierce SK, “Cloning of the gene encoding peptide-binding protein 74 shows that it is a new member of the heat shock protein 70 family”. Mol Cell Biol 1993, 13(6): 3598–610.
19. Bulut Y, Michelsen KS, Hayrapetian L, Naiki Y, Spallek R, Singh M, Arditi M, “*Mycobacterium tuberculosis* heat shock proteins use diverse toll-like receptor pathways to activate pro-inflammatory signals”, J Biol Chem 2005, 280(22): 20961-20967.
20. Fang H, Wu Y, Huang X, Wang W, Ang B, Cao X, Wan T, “Toll-like receptor 4 (TLR4) is essential for Hsp70-like protein 1 (HSP70L1) to activate dendritic cells and induce Th1 response”, J Biol Chem 2011, 286(35): 30393-30400.
21. Zeilstra-Ryalls J, Fayet O, Georgopoulos C, "The universally conserved GroE (Hsp60) chaperonins", Annu Rev Microbiol 1991, 45: 301–325.
22. Wadhwa R, Yaguchi T, Hasan MK, Mitsui Y, Reddel RR, Kaul SC, “Hsp70 family member, mot-2/mthsp70/GRP75, binds to the cytoplasmic sequestration domain of the p53 protein”, Exp Cell Res 2002, 274(2): 246-253.

23. Choi B, Choi M, Park C, Lee EK, Kang DH, Lee DJ, Yeom JY, Jung Y, Kim J, Lee S, Kang SW, "Cytosolic Hsp60 orchestrates the survival and inflammatory responses of vascular smooth muscle cells in injured aortic vessels", Cardiovasc Res 2015, 106(3): 498-508.
24. Pietrement C, Sun-Wada GH, Silva ND, McKee M, Marshansky V, Brown D, Futai M, Breton S, "Distinct expression patterns of different subunit isoforms of the V-ATPase in the rat epididymis", Biol Reprod 2006, 74(1): 185-194.
25. Lu N, Zhou Z, "Membrane trafficking and phagosome maturation during the clearance of apoptotic cells", Int Rev Cell Mol Biol 2012, 293: 269-309.
26. Vergne I, Fratti RA, Hill PJ, Chua J, Belisle J, Deretic V, "*Mycobacterium tuberculosis* phagosome maturation arrest: mycobacterial phosphatidyl-inositol analog phosphatidylinositol mannoside stimulates early endosomal fusion", Mol Biol Cell 2004, 15(2): 751-760.
27. Fratti RA, Chua J, Vergne I, Deretic V, "*Mycobacterium tuberculosis* glycosylated phosphatidylinositol causes phagosome maturation arrest", Proc Natl Acad Sci U S A 2003, 100(9): 5437-5442.
28. Luo H, Lin Y, Gao F, Zhang CT, Zhang R, "DEG 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements", Nucleic Acids Res 2014, 42(Database issue): D574-D580.
29. Lew JM, Mao C, Shukla M, Warren A, Will R, Kuznetsov D, Xenarios I, Robertson BD, Gordon SV, Schnappinger D, Cole ST, Sobral B, "Database resources for the tuberculosis community", Tuberculosis (Edinb) 2013; 93(1): 12-17.

Summary

Summary

Tuberculosis is one of the devastating diseases reported to kill a million of individuals every year. The inclination of the number of multi-drug resistant tuberculosis cases has made it mandatory for the efficient study of its causative agent, "*Mycobacterium tuberculosis*" (*MTB*). The goals of global initiatives such as STOP TB Partnership are viable to succeed only with the application of interdisciplinary research. With the rising of information data from various disciplines, the computational approach is promising to analyze and interpret the data at the system level. The study of protein interactions and the interactions network is the key element in understanding this ancient and successful pathogen. Till date, various computational methods are emerged to detect protein-protein interactions (PPIs), such as Gene neighborhood, Gene fusion, Phylogenetic profile, Mirror tree and Interlog methods. The present study has implemented these methods in *MTB* to identify a set of consistent PPIs. Further, the microarray expression data is integrated to infer 8,243 as the highly confident PPIs. Using these PPIs, a protein-interactions network is built and the topological properties are derived from it. 95 hub proteins in the core interactions are identified which could be potential drug targets for tuberculosis disease. The core interactions network has also revealed a set of 35 novel PPIs which could help in better understanding of survival and persistent mechanism of the pathogen. The novel PPIs are observed to have some virulent and conserved hypothetical proteins. Extended study of these PPIs will have a significant impact on the tuberculosis research. During infection, the pathogen evades the *Human* (*Homo sapiens*) cells and survives inside the macrophages. Living inside the macrophages, it modulates the *Human* immune response through molecular interactions for the survival and growth. Therefore, in the present study, *In silico* Two Hybrid system (I2H) is employed to predict the physical *Human-M. tuberculosis* protein interactions (*HMIs*). Interlog method along with I2H is used to identify the consistent *HMIs.* Further, the *HMIs* are integrated with functional annotations and microarray

expression data to infer 419 as the highly confident *HMIs*. A network of these confident *HMIs* is built and the highly interacting *Human* and *MTB* proteins are derived from it. It is observed that 9 *MTB* proteins are the new and significant proteins which make molecular interactions during infection. Further, the interacting proteins are examined in tuberculosis pathways to investigate *MTB* proteins that play a crucial role during the pathogenesis of tuberculosis. It is detected that tuf and gap proteins are the confident proteins and the essential gene products, which can be explored for the new therapeutic intervention strategies.

Anti-Plagiarism Report

Investigation of Protein Interaction Networks in Mycobacterium tuberculosis using Computational approaches

by Dharmapal Burne

Submission date: 17-Jan-2019 02:25PM (UTC+0530)

Submission ID: 1065166864

File name: DBharne_PhDThesisContent_Plag.pdf (744.41K)

Word count: 14786

Character count: 80452

Investigation of Protein Interaction Networks in Mycobacterium tuberculosis using Computational approaches

ORIGINALITY REPORT

4%	3%	2%	0%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	bif.uohyd.ac.in Internet Source	2%
2	Iñaki Comas. "The Past and Future of Tuberculosis Research", PLoS Pathogens, 10/26/2009 Publication	<1%
3	Gomez, J.E.. "M. tuberculosis persistence, latency, and drug tolerance", Tuberculosis, 2004 Publication	<1%
4	en.wikipedia.org Internet Source	<1%
5	Dyer, M.D.. "Supervised learning and prediction of physical interactions between human and HIV proteins", Infection, Genetics and Evolution, 201107 Publication	<1%
6	theses.ucalgary.ca Internet Source	<1%

7	dukespace.lib.duke.edu Internet Source	<1%
8	ir.library.osaka-u.ac.jp Internet Source	<1%
9	Peipei Li, Lyong Heo, Meijing Li, Keun Ho Ryu, Gouchol Pok. "Protein function prediction using frequent patterns in protein-protein interaction networks", 2011 Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), 2011 Publication	<1%
10	physiolgenomics.physiology.org Internet Source	<1%
11	Rana, Amrita K., Albel Singh, Sudagar S. Gurcha, Liam R. Cox, Apoorva Bhatt, and Gurdyal S. Besra. "Ppm1-Encoded Polyprenyl Monophosphomannose Synthase Activity Is Essential for Lipoglycan Synthesis and Survival in Mycobacteria", PLoS ONE, 2012. Publication	<1%
12	Joao Leitao, Jose Pereira, Luis Rodrigues. "HyParView: A Membership Protocol for Reliable Gossip-Based Broadcast", 37th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN'07), 2007 Publication	<1%

13 Adam, Nathalie, Lucia Vergauwen, Ronny Blust, and Dries Knapen. "Gene transcription patterns and energy reserves in Daphnia magna show no nanoparticle specific toxicity when exposed to ZnO and CuO nanoparticles.", Environmental Research, 2015.
Publication
<1%

14 XiaoZhen Wang. "Genes and regulatory networks involved in persistence of Mycobacterium tuberculosis", Science China Life Sciences, 01/21/2011
Publication
<1%

Exclude quotes	On	Exclude matches	< 14 words
Exclude bibliography	On		

PUBLICATIONS

Life Science Informatics Publications

Research Journal of Life Sciences, Bioinformatics, Pharmaceutical and Chemical Sciences

Journal Home page http://www.rjlbpcs.com/

RJLBPCS
ISSN 2454-6348

Original Research Article

DOI: 10.26479/2018.0406.05

INFERRING PROTEIN INTERACTION NETWORK OF *MYCOBACTERIUM TUBERCULOSIS H37RV* USING SEQUENCE INFORMATION

Dhammapal Bharne, Damuka Naresh, Vaibhav Vindal*

Department of Biotechnology and Bioinformatics, School of Life Sciences,
University of Hyderabad, Hyderabad, India.

ABSTRACT: Protein interaction network helps to understand the general mechanism behind the complex biological systems which in turn provide insights into disease mechanism and pathways. It can be modulated towards drug discovery for the infectious organism. With vast sequence data availability, computational methods play a major role to infer protein-protein interaction. The present study explores sequence-based information to infer protein interaction network in *Mycobacterium tuberculosis*, a causative agent of one of the leading infectious diseases Tuberculosis. A model was built using support vector machine (SVM) based classification through amino acid conjoint triad from interacting protein pairs from the DIP database. 162,528 protein interaction pairs of *M. tuberculosis H37Rv* were identified through the model. It was observed that 53,602 protein pairs were already known to be interacting or co-expressing in multiple microarray experiments. These protein pairs were considered as significant interaction pairs. A protein interaction network built for core interactions had 53,424 edges connected across 1,368 nodes. The highest degree was observed for pknB, a serine/threonine-protein kinase which may serve as a potential drug target for tuberculosis. Further, two conserved hypothetical proteins, Rv3879c and Rv3909, were found to be hub proteins. Exploration of such hubs will assist in understanding the regulation of genes and disease processes which may lead to develop better intervention strategies for the disease.

KEYWORDS: Support vector machine, conjoint triad, protein interaction network, degree, hub.

Corresponding Author: Dr. Vaibhav Vindal* Ph.D.

Department of Biotechnology and Bioinformatics, School of Life Sciences,
University of Hyderabad, India. Email Address: vaibhav@uohyd.ac.in

1. INTRODUCTION

Now is the time for modern biology to switch from the study of single molecules to network-based systems. The functions of various cellular processes are determined by protein interactions rather than an individual protein. Though, a large number of protein-protein interactions (PPIs) are predicted with experimental methods, it covers only a fraction of total protein-protein interactions [1, 2]. In this regard, computational methods play a vital role in achieving the complete protein interaction network [3, 4, 5]. Various in silico methods are proposed to identify PPIs such as phylogenetic profile, gene operon and domain-based methods [6, 7, 8]. Recently, it is reported that the identification of PPIs based on only sequence information is more universal [9, 10]. Amino acid conjoint triad method identifies PPIs using only protein sequence information [10]. The sequence based method is powerful for identifying protein-protein interactions and in exploring the networks for newly discovered proteins with unknown biological functions. In spite of large research efforts, tuberculosis is still one of the leading infectious diseases in the world [11, 12]. Multidrug resistance and HIV co-infections provoke to investigate a novel system to combat the disease [13]. A protein interaction network of *M. tuberculosis H37Rv* helps in understanding cellular physiology and also identifying suitable drug targets [14]. The present study employs amino acid conjoint triad method to identify interacting protein pairs in *M. tuberculosis H37Rv*. Since the interaction network is a large scale real world graphical data, it is an ideal challenge for bioinformatics research.

2. MATERIALS AND METHODS

1. Dataset Preparation

Protein interaction data for *M. tuberculosis H37Rv* was downloaded from DIP databases [15]. These interacting pairs were used to prepare positive and negative training examples for conjoint triad method [10]. During this process, 343 triads of amino acids with similar physiochemical roles were generated. The frequency of each triad was calculated and then normalized to represent a protein sequence. An interaction pair was obtained by concatenating two proteins with normalized triad frequencies. Therefore, each interaction pair had 686 conjoint triads of amino acids. Since the protein interactions are symmetrical, a reverse directional calculation was also employed. In order to prepare negative training examples, Euclidean distances among the proteins from the interaction pairs were calculated. Protein pairs with Euclidean distance value above the average of smallest and largest Euclidean distance values were used as negative training examples.

2. Generation of SVM Model

In order to generate a SVM model, radial basis function (RBF) from LIBSVM software [16] was employed. Positive and negative training examples were scaled together with default parameters. The scaled data was then randomized 1000 times and grid search was performed to identify the best C and γ parameter values for the RBF kernel. The scaled data was trained using the best C and γ values to generate a SVM model.

3. Prediction of PPIs

Complete proteome set of *M. tuberculosis H37Rv* was downloaded from NCBI genome [https://www.ncbi.nlm.nih.gov/genome/]. Conjoint triad frequencies were calculated and normalized for each of the protein sequences. Protein pairs were generated by concatenating every protein sequence with every other protein sequence. If a protein pair was predicted positive by the built SVM model, it was considered as interacting pairs.

4. Inferring Protein interaction map

Predicted protein interaction pairs were compared with interaction data from the STRING database [17, 18] and the MPIDB database [19]. They were also compared with already known interactions from Wang *et al*., [20] and Liu *et al*., [21] data. Further, predicted interaction pairs were validated using co-expression analysis [22]. During this process, microarray experiments from NCBI GEO [https://www.ncbi.nlm.nih.gov/geo/] were filtered with "*Mycobacterium tuberculosis*" and a sample size of at least 10. Further, only experiments with ORF names were considered. The experimental data was log base 2 transformed and imputed with k-nearest neighbor method. It was quantile normalized to remove sources of variations. Expression profiles of the proteins in the predicted protein interactions were extracted from the normalized data and Pearson correlation coefficients were calculated. The pairs of proteins with correlation value of at least 0.8 in more than one microarray experiment were considered as co-expressing proteins pairs [22]. Predicted protein pairs supported by interaction data from the STRING database, Wang *et al.,* data, Liu *et al.,* data or co-expressing in multiple microarray experiments were considered as significant protein interaction pairs. An interaction map was generated for these significant pairs using VisANT software [23, 24]. High degree nodes and clustering coefficients were derived from the interaction network.

5. Functional Enrichment

Top 5 percent high degree nodes of the protein interaction network were considered as hubs. These hubs were used to find their functional enrichment in ontologies such as biological process, cellular component and molecular function through PANTHER over-representation test [25]. P-value cut off is set to 0.05 in order to obtained significant ontological terms. Bonferroni correction for multiple testing was also considered during the analysis.

3. RESULTS AND DISCUSSION

1. Protein-protein interactions

It was observed that the training data set had 38 positive examples and 444 negative examples. The grid search of the scaled data generated the best C value of 8.0 and γ value of 0.0078125. A SVM model generated using these values through RBF kernel was used to predict protein interactions in *M. tuberculosis H37Rv*. In the present study, a total of 162,528 PPIs were predicted using the built SVM model. It was observed that the predicted interactions were scattered among 1766 proteins. However, it leads to the coverage of 45.12 percent of total proteins of *M. tuberculosis H37Rv*.

2. Known Interacting Pairs

It was observed that some of the predicted PPIs were already reported in the existing literature such as Wang et al. and Liu et al. data. Some of them were also available in the databases such as STRING database and MPIDB database. Further, 75 microarray experiments were found to have ORF names in expression data. Co-expression analysis of microarray experiments for the proteins in the predicted PPIs was also performed. The Table 1 represents the number of predicted PPIs that overlap with existing literature and databases as well as co-expresses in multiple microarray experiments. It is clear from the table that 59,019 PPIs overlaps with existing knowledge. It was found to have 53,602 unique PPIs. Therefore, more than 32.98 percent of the predicted PPIs were supported from the know data. These predicted PPIs were considered as significant proteins interaction pairs in the present study.

Table 1: Overlapping PPIs

Sr. No.	Source	Dataset Size	Overlapped PPIs
1	STRING database	796,610	15,433
2	MPIDB database	19	5
3	Liu et al.	43,136	1,304
4	Wang et al.	8,242	60
5	Co-expression analysis	1,705,422	42,217
		Total PPIs	**59,019**

3. Protein Interaction Network

Significant protein interaction pairs were used to construct a protein interaction map of *M. tuberculosis H37Rv*. It was observed that there were 183 isolated nodes with a degree of 1. These nodes were removed to get a core interaction network. The Figure 1 represents an interaction map generated from the core interaction network using the VisANT software. The network has 53,424 black colored lines called edges representing core protein interactions and 1,368 green colored boxes called nodes representing the interacting proteins. Visualization of the interaction network indicates that it is a simple network. Degree distribution follows the power law; hence the interaction network is a scale-free network [26]. The highest degree was observed for Rv0014 (pknB) with the value of 871 followed by Rv2524c (fas) with the value of 663 and Rv2379c (mbtF) with the value of 648. When the degree threshold was set at 50 and above, 608 hub nodes were identified. Correlation of node degree with clustering coefficient was found to be 0.44 indicating that it is a well clustered network. The average of clustering coefficients of the nodes was found to

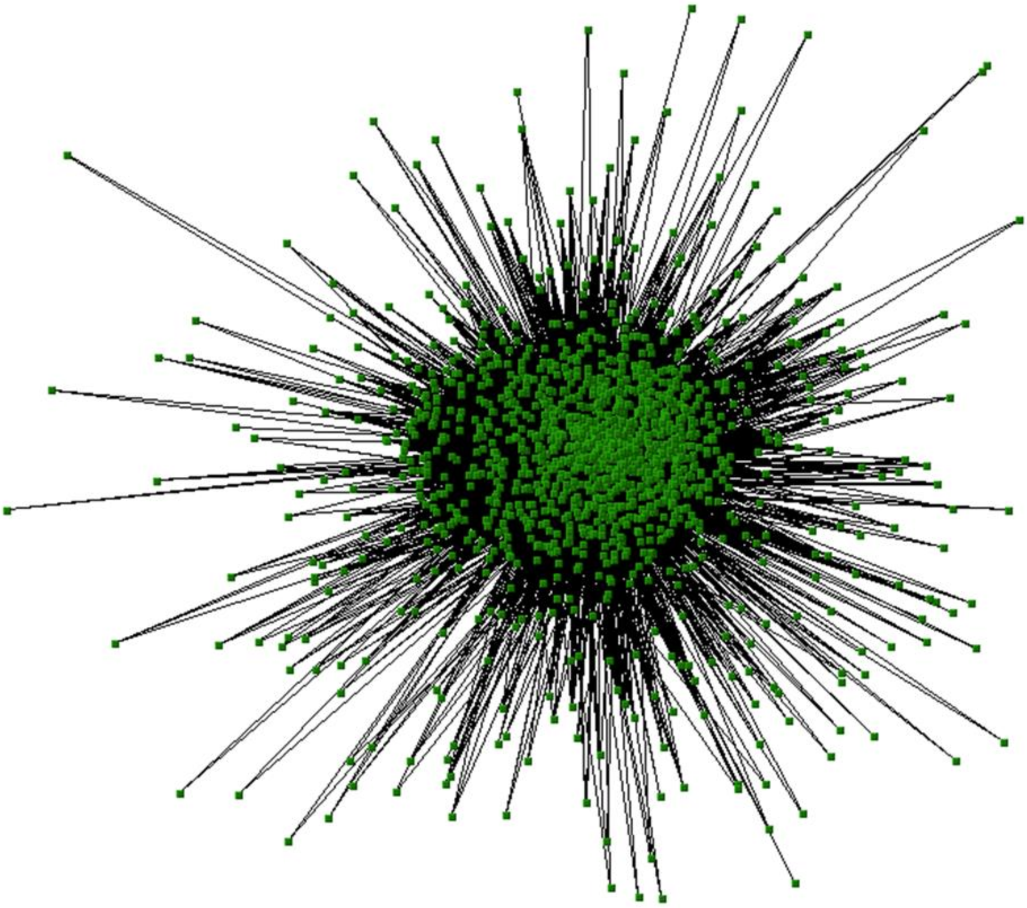


Figure 1: Protein Interaction network of *M. tuberculosis H37Rv*

be 0.459 suggesting that there are many nodes which were well clustered. There were 30 fully connected subgraphs. These subgraphs represent related functions of the protein and strongly suggest that they could form a complex [27]. As the network is a scale-free network, deletion of high degree nodes will destroy the interaction network significantly. Therefore, they can be exploited as potential targets for effective control and cure of the disease.

4. Functional annotation

Top 5 percent of hubs were employed for functional annotation [28, 29]. It was observed the cutoff degree to be 378 which was possessed by three nodes, viz., Rv2447, Rv3879c and Rv2931. Thus value 5 percent of hubs constituted 33 nodes. Functional annotation of these nodes was performed using PANTHER [16] over-representation test. It indicated that most of the hub proteins were involved in fatty acids, amino acids and lipid metabolic processes. Further, they were mostly enriched in transferase, hydrolase and ligase activities.

5. Functional categories

Top 5 percent hub proteins were categories based on Tuberculist [30] data. The Figure 2 indicates the proportion of hubs under different categories. It is clear from the figure that most of the hubs are involved in lipid metabolism, cell wall and cell processes. Further, it was observed that Rv3879c and Rv3909 the two conserved hypothetical proteins which are also hubs. Further, it was observed that 16 hubs are essential genes as observed from the Database of Essential Genes [31, 32].

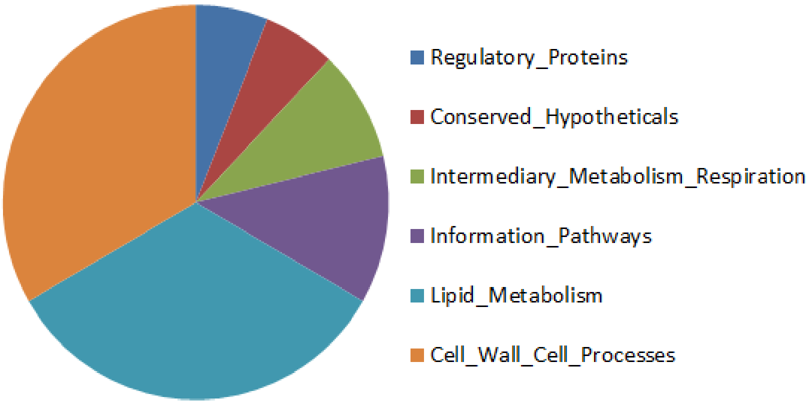


Figure 2: Tuberculist categories of Hubs

4. CONCLUSION

In the present study, support vector machine based conjoint triad was efficiently employed to predict PPIs of *M. tuberculosis H37Rv*. Using this approach, 162,528 interacting pairs were identified. It was observed that more than 32.98 percent of predicted PPIs overlap with already known PPIs. Therefore, this approach is effective to identify PPIs using only sequence information. The proteome coverage of predicted PPIs was 45.12. This suggests that different approaches can be employed to predict number of interacting pairs in order to achieve a complete protein network of *M. tuberculosis H37Rv*. Hubs were identified which may serve as potential drug targets for tuberculosis. Further, conserved hub hypothetical proteins were also found which can be explored further to understand their potential role in cellular physiology and disease mechanism of the organism.

ACKNOWLEDGEMENT

The Departmental facilities were supported by grants from UPE-II and DST-PURSE. DB is supported by UGC fellowship. Authors acknowledge computational support from Bioinformatics Infrastructure Facility, University of Hyderabad, India.

CONFLICT OF INTEREST

None.

REFERENCES

1. Zhou H, Wong L. Comparative analysis and assessment of *M. tuberculosis H37Rv* protein-protein interaction datasets. BMC Genomics. 2011; 12(3):S20.
2. Han J-DJ, Dupuy D, Bertin N, Cusick ME, Vidal M. Effect of sampling on topology predictions of protein-protein interaction networks. Nat Biotechnol. 2005; 23(7):839-844.
3. Wodak SJ, Méndez R. Prediction of protein–protein interactions: the CAPRI experiment, its evaluation and implications. Curr Opin Struct Biol. 2004; 14(2):242-249.
4. Valencia A, Pazos F. Computational methods for the prediction of protein interactions. Curr Opin Struct Biol. 2002; 12(3):368-373.
5. Raman K. Construction and analysis of protein–protein interaction networks. Autom Exp. 2010;

2(1):2.

6. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci U S A. 1999; 96(8):4285-4288.
7. Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA. Protein interaction maps for complete genomes based on gene fusion events. Nature. 1999; 402(6757):86-90.
8. Jothi R, Cherukuri PF, Tasneem A, Przytycka TM. Co-evolutionary analysis of domains in interacting proteins reveals insights into domain–domain interactions mediating protein–protein interactions. J Mol Biol. 2006; 362(4):861-875.
9. Skrabanek L, Saini HK, Bader GD, Enright AJ. Computational prediction of protein–protein interactions. Mol Biotechnol. 2008; 38(1):1-17.
10. Shen J, Zhang J, Luo X, Zhu W, Yu K, Chen K, et al. Predicting protein–protein interactions based only on sequences information. Proc Natl Acad Sci U S A. 2007; 104(11):4337-4341.
11. Hingley-Wilson SM, Sambandamurthy VK, Jacobs Jr WR. Survival perspectives from the world's most successful pathogen, *Mycobacterium tuberculosis*. Nat Immunol. 2003; 4(10):949-955.
12. Raviglione M, Sulis G. Tuberculosis 2015: burden, challenges and strategy for control and elimination. Infect Dis Rep. 2016; 8(2):6570.
13. Leibert E, Danckers M, Rom WN. New drugs to treat multidrug-resistant tuberculosis: the case for bedaquiline. Ther Clin Risk Manag. 2014; 10:597-602.
14. 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.
15. Xenarios I, Salwinski L, Duan XJ, Higney P, Kim S-M, Eisenberg D. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. Nucleic Acids Res. 2002; 30(1):303-305.
16. Chang CC, Lin CJ. LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. 2011; 2(3):27.
17. Mering Cv, Huynen M, Jaeggi D, Schmidt S, Bork P, Snel B. STRING: a database of predicted functional associations between proteins. Nucleic Acids Res. 2003; 31(1):258-261.
18. Von Mering C, Jensen LJ, Snel B, Hooper SD, Krupp M, Foglierini M, et al. STRING: known and predicted protein–protein associations, integrated and transferred across organisms. Nucleic Acids Res. 2005; 33:D433-D437.
19. Goll J, Rajagopala SV, Shiau SC, Wu H, Lamb BT, Uetz P. MPIDB: the microbial protein interaction database. Bioinformatics. 2008; 24(15):1743-1744.
20. Wang Y, Cui T, Zhang C, Yang M, Huang Y, Li W, et al. Global protein- protein interaction

network in the human pathogen *Mycobacterium tuberculosis H37Rv*. J Proteome Res. 2010; 9(12):6665-6677.

21. Liu ZP, Wang J, Qiu YQ, Leung RK, Zhang XS, Tsui SK, et al. Inferring a protein interaction map of *Mycobacterium tuberculosis* based on sequences and interologs. BMC Bioinformatics; 2012; 13:S6.
22. Lee HK, Hsu AK, Sajdak J, Qin J, Pavlidis P. Coexpression analysis of human genes across many microarray data sets. Genome Res. 2004; 14(6):1085-1094.
23. Hu Z, Mellor J, Wu J, DeLisi C. VisANT: an online visualization and analysis tool for biological interaction data. BMC Bioinformatics. 2004; 5(1):17.
24. Granger BR, Chang YC, Wang Y, DeLisi C, Segre D, Hu Z. Visualization of metabolic interaction networks in microbial communities using VisANT 5.0. PLoS Comput Biol. 2016; 12(4):e1004875.
25. Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, et al. PANTHER: a library of protein families and subfamilies indexed by function. Genome Res. 2003; 13(9):2129-2141.
26. Barabasi AL, Albert R. Emergence of scaling in random networks. Science. 1999; 286(5439):509-512.
27. Langfelder P, Horvath S. Fast R functions for robust correlations and hierarchical clustering. J Stat Softw. 2012; 46(11):i11.
28. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nat Genetics. 2000; 25(1):25-29.
29. Mi H, Poudel S, Muruganujan A, Casagrande JT, Thomas PD. PANTHER version 10: expanded protein families and functions, and analysis tools. Nucleic Acids Res. 2015; 44(D1):D336-D342.
30. Lew JM, Mao C, Shukla M, Warren A, Will R, Kuznetsov D, et al. Database resources for the tuberculosis community. Tuberculosis (Edinb). 2013; 93(1):12-17.
31. Zhang R, Ou HY, Zhang CT. DEG: a database of essential genes. Nucleic Acids Res. 2004; 32:D271-D272.
32. Luo H, Lin Y, Gao F, Zhang CT, Zhang R. DEG 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements. Nucleic Acids Res. 2013; 42(D1):D574-D580.

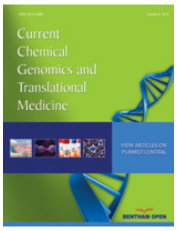
BENTHAM OPEN

CrossMark

Current Chemical Genomics and Translational Medicine

Content list available at: www.benthamopen.com/CCGTM/

DOI: 10.2174/221398850181201????, 2018, *12*, 03-00



LETTER

maGUI: A graphical User Interface for Analysis and Annotation of DNA Microarray Data

Dhammapal Bharne, Praveen Kant[$] and Vaibhav Vindal[*]

Department of Biotechnology and Bioinformatics, School of Life Sciences, University of Hyderabad, Hyderabad, India
[$]Present Address: *Department of Biosciences and Bioengineering, Indian Institute of Technology Bombay, Mumbai, India*

Received: September 7, 2018 Revised: October 29, 2018 Accepted: October 30, 2018

Abstract:

Summary:

maGUI is a graphical user interface designed to analyze microarray data produced from experiments performed on various platforms such as Affymetrix, Agilent, Illumina, and Nimblegen and so on, automatically. It follows an integrated workflow for pre-processing and analysis of the microarray data. The user may proceed from loading of microarray data to normalization, quality check, filtering, differential gene expression, principal component analysis, clustering and classification. It also provides miscellaneous applications such as gene set test and enrichment analysis and identifying gene symbols using Bioconductor packages. Further, the user can build a co-expression network for differentially expressed genes. Tables and figures generated during the analysis can be viewed and exported to local disks. The graphical user interface is very friendly especially for the biologists to perform the most microarray data analyses and annotations without much need of learning R command line programming.

Availability and Implementation:

maGUI is an R package which can be downloaded freely from Comprehensive R Archive Network resource. It can be installed in any R environment with version 3.0.2 or above.

Keywords: : Graphical user interface, R programming language, Bioconductor, Comprehensive R Archive Network, Microarray data analysis, Gene set test analysis, Gene set enrichment analysis.

1. INTRODUCTION

A large number of experiments are carried out on microarrays developed by different manufacturers such as Affymetrix, Agilent, Illumina and Nimblegen and so on. The data generated from these experiments are in various formats and hence the pre-processing is different. Though R programming language [1] provides the most sophisticated software environment to analyze and annotate the microarray data, it is difficult for the biologists who are not familiar with the programming languages. Further, the available packages and software are restricted to either a single platform or to the extent of analysis, and also the manual processing and analysis of the data is a time-consuming and tedious process, it is extremely important to develop an R package that integrates and simplifies the process of analysis and annotation of microarray data belonging to different platforms.

In the present study, “maGUI”, a graphical user interface (GUI) [2], is developed using gWidgets, tcltk and other packages of R to analyze and annotate the microarray data easily and more user friendly. It integrates limma [3], affy [4], lumi [5, 6] and several other packages of Bioconductor [7]. It enables the pre-processing of microarray data and identification of differentially expressed genes automatically. Further, the user can identify functional categories and pathways for different genes in the microarray data. Therefore, the GUI is very much useful in solving challenges arising during the analysis and annotation of DNA microarray data.

[*] Address correspondence to this author at the Department of Biotechnology and Bioinformatics, School of Life Sciences, University of Hyderabad, Hyderabad, India, Tel: 23134589; E-mail: vaibhav@uohyd.ac.in

2. APPROACH

2.1. The maGUI User Interface

The maGUI package is available at the Comprehensive R Archive Network (CRAN) repository and hence the GUI can be installed as any other R package. It provides the user with a graphical user interface on top of the normal graphical functions, allowing the user to interactively pre-process and analyze the microarray data easily and efficiently. The GUI consists of menus for pre-processing, analyzing, and annotating the microarray data, a container for hierarchical nature of tasks performed on any microarray data and a graphical region for viewing figures and tables that are generated during analysis and annotation of the microarray data. The following figure, Fig. (**1**), represents a flow chart for the application of maGUI.

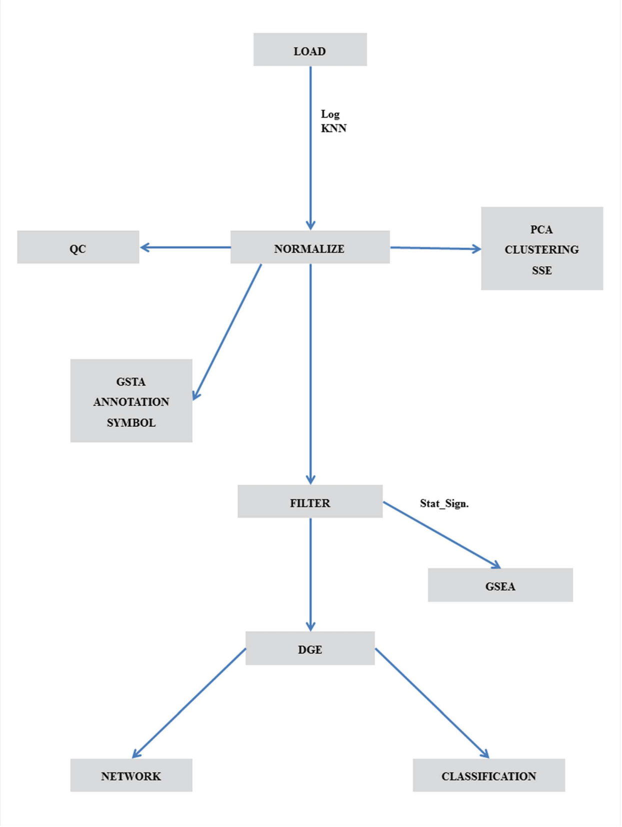


Fig. (1). Work flow and application of maGUI.

2.2. Pre-Processing Microarray Data

Once the GUI is launched, the user can import the microarray data of various platforms from the File menu. The user can also load microarray SOFT file and series matrix file for the analysis. The loaded microarray data can be preprocessed and normalized from Preprocessing menu. The GUI utilizes different packages for quantile normalization such as limma, bead array [8] or lumi based on the microarray experimental platform. Quality assessment is made using quality control plots and box plots. Such data is used for clustering, classification, finding differentially expressed genes and so on from the Analysis menu.

2.3. Analysis of Microarray Data

maGUI facilitates clustering of samples with Pearson correlation coefficient and complete linkage methods. It plots principal component analysis (PCA) using the singular value decomposition method. The normalized data is filtered specifically providing control and test sample names of one or multiple groups. Alternatively, the user can choose for unspecific filtering with expression filter or standard deviation filter. Both specific and unspecific filtering employs fitting linear model and empirical Bayes moderation provided by limma package. During filtering of data, genes with p-value less than 0.01 are extracted and stored as statistically significant genes. Differentially expressed genes (DGE) are identified based on the number of groups in specific filtering or using unspecific filtered data. Further, the user can filter out the top differentially expressed genes with the log fold change value. Classification of data is performed based on the expression profiles of differentially expressed genes and can be viewed as red and green color heat map. Such an analysis helps the user in not only describing the relationships between genes but also characterizing the specific molecular differences associated with them [9]. The following figure Fig. (**2**) is a heatmap of top 10 differentially expressed genes for the NCBI GEO [10] experiment number GSE6141, which was performed for the global analysis of the *Drosophila* NELF complex [11].

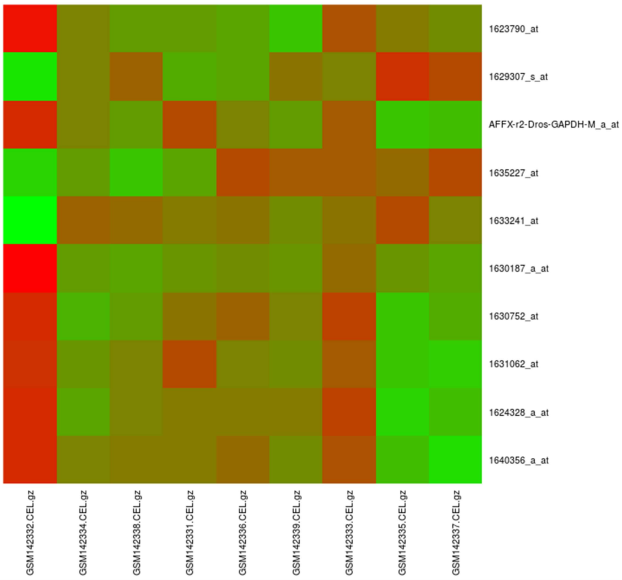


Fig. (2). Heatmap of top 10 differentially expressed genes. Samples GSM142332.CEL.gz, GSM142331.CEL.gz, and GSM142333.CEL.gz are the controls while GSM142334.CEL.gz, GSM142336.CEL.gz and GSM142335.CEL.gz are LacZ treated samples and GSM142338.CEL.gz, GSM142339.CEL.gz and GSM142337.CEL.gz are NELF depleted samples.

2.4. Gene Set Analysis

Expression data of the significant genes are combined with other knowledge to find the functional relevance of the genes. This is achieved through the enrichment of the genes under different biochemical pathways [12] and functional categories such as biological processes, molecular functions and cellular components [13]. In addition, all genes of the microarray data can be annotated to various GO domains and KEGG pathways through gene set test analysis. Both the gene set enrichment and test analyses can be performed from the Miscellaneous menu of the maGUI utilizing annotation databases [14, 15]. The source for these annotation databases is the Bioconductor [7], which is an open source and open development. With the maGUI package, the GO terms of any domain can be generated as graphs with yellow nodes representing the genes present in the current microarray data while white nodes as their parents. The KEGG pathways can also be generated as a graph using their KEGG Ids and inbuilt organism codes obtained from KEGG resource. In the KEGG graph, red-colored nodes represent up-regulated genes while green colored nodes represent down-regulated genes. Nodes colored in grey are the genes present but are not differentially regulated in the current microarray data while white colored nodes are their parents. Thus, the maGUI helps not only in identifying gene regulations and pathways but also in making interesting biological interpretations from the microarray data [16].

2.5. Additional Features

The maGUI enables the user to perform and visualize sample size estimation with 2 fold change which is critical in designing any microarray experiment. It also maps all the identifiers from normalized microarray data to their corresponding gene symbols using the annotation database. Further, it builds a co-expression network using expression correlation of differentially expressed genes [17]. A network of such links helps in identifying genes associated with a disease state. The user can also identify protein-protein associations among all the genes in two different normalized microarray data using the correlation of co-expression profile of each gene [18]. Such protein associations play a major role in identifying various cellular and biochemical pathways.

2.6. Data Export

All the tables and figures generated during microarray data analysis such as clustering of samples, PCA, classification and so on can be visualized from the View menu and exported to local drives from the Export menu. Further, the images can be directly saved from the graphical region. Objects generated during microarray data analysis and annotation can be saved as an R data file.

2.7. Availability and Implementation

maGUI is an R package that can be freely downloaded from CRAN resource. It is associated with various other packages such as gWidgets, RGtk2, RSQLite and so on which will be installed along with the package in any R environment with version 3.0.2 or later. It is successfully tested on Linux, Windows and OS X operating systems. The reference manual of the package is available at https://cran.r-project.org/web/packages/maGUI/maGUI.pdf. Tutorial documentation with examples for various applications of the maGUI can be downloaded from http://bif.uohyd.ac.in/maGUI/maGUI_Tutorial.pdf.

CONCLUSION

maGUI is a user-friendly, cross-platform GUI for analysis and annotation of microarray data. It provides various features for efficient analysis and interpretation of the microarray data. It also relates genes to various knowledge-based databases to infer functional significance. The GUI is especially useful for the biologists who are not familiar with any programming language. The package is freely available at CRAN resource (https://cran.r-project.org/web/packages/maGUI/).

CONSENT FOR PUBLICATION

Not applicable.

CONFLICT OF INTEREST

The authors declare no conflict of interest, financial or otherwise.

ACKNOWLEDGEMENTS

Authors would like to thank funding support from DST-PURSE, UPE to Department of Biotechnology and Bioinformatics, University of Hyderabad. DB is supported by UGC fellowship. Authors also thank the Bioinformatics Infrastructure Facility, University of Hyderabad for hosting supportive files.

REFERENCES

[1] Core R , Team A. Language and Environment for Statistical Computing R Foundation for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing 2013.

[2] Lawrence M, Verzani J. Programming Graphical User Interfaces in R. Chapman and Hall/CRC The R Series, 2012.

[3] Ritchie ME, Phipson B, Wu D, *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 2015; 43(7): e47.
[http://dx.doi.org/10.1093/nar/gkv007] [PMID: 25605792]

[4] Gautier L, Cope L, Bolstad BM, Irizarry RA. affy--analysis of Affymetrix GeneChip data at the probe level. Bioinformatics 2004; 20(3): 307-15.
[http://dx.doi.org/10.1093/bioinformatics/btg405] [PMID: 14960456]

[5] Lin SM, Du P, Huber W, Kibbe WA. Model-based variance-stabilizing transformation for Illumina microarray data. Nucleic Acids Res 2008; 36(2): e11.
[http://dx.doi.org/10.1093/nar/gkm1075] [PMID: 18178591]

[6] Du P, Kibbe WA, Lin SM. lumi: A pipeline for processing Illumina microarray. Bioinformatics 2008; 24(13): 1547-8.
[http://dx.doi.org/10.1093/bioinformatics/btn224] [PMID: 18467348]

[7] Gentleman RC, Carey VJ, Bates DM, *et al.* Bioconductor: Open software development for computational biology and bioinformatics. Genome Biol 2004; 5(10): R80.
[http://dx.doi.org/10.1186/gb-2004-5-10-r80] [PMID: 15461798]

[8] Dunning MJ, Smith ML, Ritchie ME, Tavaré S. beadarray: R classes and methods for Illumina bead-based data. Bioinformatics 2007; 23(16): 2183-4.
[http://dx.doi.org/10.1093/bioinformatics/btm311] [PMID: 17586828]

[9] Planet PJ, DeSalle R, Siddall M, Bael T, Sarkar IN, Stanley SE. Systematic analysis of DNA microarray data: Ordering and interpreting

patterns of gene expression. Genome Res 2001; 11(7): 1149-55.
[http://dx.doi.org/10.1101/gr.187601] [PMID: 11435396]

[10] Barrett T, Wilhite SE, Ledoux P, *et al.* NCBI GEO: Archive for functional genomics data sets--update. Nucleic Acids Res 2013; 41(Database issue): D991-5.
[PMID: 23193258]

[11] Gilchrist DA, Nechaev S, Lee C, *et al.* NELF-mediated stalling of Pol II can enhance gene expression by blocking promoter-proximal nucleosome assembly. Genes Dev 2008; 22(14): 1921-33.
[http://dx.doi.org/10.1101/gad.1643208] [PMID: 18628398]

[12] Minoru K, Susumu G. Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 2000; 28(1): 27-30.

[13] Ashburner M, Ball CA, Blake JA, *et al.* Gene ontology: Tool for the unification of biology. Nat Genet 2000; 25(1): 25-9.
[http://dx.doi.org/10.1038/75556] [PMID: 10802651]

[14] Zhu Y, Davis S, Stephens R, Meltzer PS, Chen Y. GEOmetadb: Powerful alternative search engine for the gene expression omnibus. Bioinformatics 2008; 24(23): 2798-800.
[http://dx.doi.org/10.1093/bioinformatics/btn520] [PMID: 18842599]

[15] Falcon S, Gentleman R. Using GOstats to test gene lists for GO term association. Bioinformatics 2007; 23(2): 257-8.
[http://dx.doi.org/10.1093/bioinformatics/btl567] [PMID: 17098774]

[16] Zhang YH, Chu C, Wang S, *et al.* The use of gene ontology term and KEGG pathway enrichment for analysis of drug half-Life. PLoS One 2016; 11(10): e0165496.
[http://dx.doi.org/10.1371/journal.pone.0165496] [PMID: 27780226]

[17] Langfelder P, Horvath S. WGCNA: An R package for weighted correlation network analysis. BMC Bioinformatics 2008; 9: 559.
[http://dx.doi.org/10.1186/1471-2105-9-559] [PMID: 19114008]

[18] Lee HK, Hsu AK, Sajdak J, Qin J, Pavlidis P. Coexpression analysis of human genes across many microarray data sets. Genome Res 2004; 14(6): 1085-94.
[http://dx.doi.org/10.1101/gr.1910904] [PMID: 15173114]