# An Integrative Approach to Studying Protein-Protein Interactions Within the Genome of *Mycobacterium tuberculosis*

Kenneth Babu Opap

· April 2011·

Department of Molecular and Cell Biology
University of Cape Town

*Submitted in partial fulfilment of the requirements for the degree of Master of Science*

Supervised by
Prof. Nicola Mulder



This work was supported by a grant from the National Bioinformatics Network, South Africa through the Computational Biology Group at the University of Cape Town.

**Abstract**

Proteins mediate bio-chemical processes in cellular organisms by functioning as structural components, signaling molecules, enzymes, transcription factors and receptors. They perform these functions through complex interactions that they make with one another. The study of protein-protein interactions (PPI) within an organism is therefore a vital step towards understanding the cellular life process of an organism. Conventionally, PPIs have been studied by specially designed experiments to identify pairs of proteins that are suspected to interact. Experimental approaches are quite expensive, yet still there are cases of wrongly identified protein-protein interactions.

In this study, we predict PPIs computationally by exploiting known properties of proteins and PPIs such as, protein domains being the agents that mediate PPIs, the ability to transfer PPIs across phylogenetically related organisms (orthologs), and the requirement that proteins be present in the same cell compartment at the same time in order to interact (subcellular localization and gene expression).

We then develop a scoring mechanism that weights methods differently according to the confidence that we attach to the plausibility of the interactions that they predict, in addition to taking into account the number of evidences that support a particular interaction, such that interactions that are supported by many methods have higher scores.

Thereafter, we elucidate the evolutionary dynamics of our set of predicted PPIs by calculating the rate of non-synonymous substitutions to that of synonymous substitutions (dN/dS) and codon volatility values with the aim of identifying whether interacting proteins co-evolve.

Finally, we evaluate the PPIs in our predicted set for biological relevance by calculating functional similarity based on GO annotation for each PPI to find out whether interacting proteins more often than not are annotated to similar GO terms. We performed GO enrichment analysis with the aim to find over-represented GO terms in set of predicted PPIs.

# Contents

# List of Figures

# List of Tables

# Acknowledgements

I thank Prof. Nicola Mulder for supervising the research. She was very supportive throughout the research project, often going out to great lengths to help me secure study materials that I required. I also acknowledge her invaluable guidance on the art of scientific writing, having read many drafts of publications and scientific presentation materials that culminated into this thesis. I also owe the accomplishment of this work partly to the members of Computational Biology Group at the University of Cape Town including, Gaston Mazandu; who helped me a great deal with the mathematical analysis, Gerrit Botha, Natasha Woods, Ayton Meintjes, Renaud Gaujoux, Gustavo Salazaar, Miguel Lacerda, Bonginkosi Mntungwa, Jean Michel Safari, Emile Chimusa, Victoria Nembaware, Tumi Mofolo, Yves Semegni, Elizabeth Kelly, Cashifa Karriem, Dr Darren Martin and Rodger Duffet, to whose technical skills, the lab owes the stable compute environment long after he left. I also specially thank Dr Wayne Delport and the HyPhy support team, who helped me set up various phylogenetic analysis programs, even after Wayne had relocated to the United States. I also acknowledge my friend Fredrick Kamanu who helped me settle in Cape Town and made the stay enjoyable. Finally, I also thank my family back in Kenya for their encouragement, support and patience for the duration of my study.

Cape Town, April 2011

# Chapter 1

# Introduction



One third of the world's population is thought to be infected with *Mycobacterium tuberculosis* (MTB), a bacterium that causes Tuberculosis or TB (short for tubercles bacillus) in humans Jasmer et al. (2002). The World Health Organisation (WHO), estimates that in the year 2007, 9.27 million new cases of TB occurred (139 individual per 100 000 population) Jasmer et al. (2002). The WHO also estimates that someone in the world is newly infected with TB bacilli every second WHO (2009), and that the mortality rate due to TB in the year 2008 was 1.3 million people. The human health devastation attributed to TB has prompted research on the *Mycobacterium tuberculosis* by researchers from diverse backgrounds. Many research studies have been targeted at finding effective means for diagnosing and treating TB.

The WHO launched the Stop TB strategy in the year 2006 in a bid to curb the devastation caused by TB. The Stop TB strategy sets out major interventions that need to be implemented in order to effectively manage TB. The interventions are divided into six broad components: (i) pursuing high quality Directly Observed Therapy, Short-course (DOTS); (ii) addressing TB/HIV, Multi-Drug Resistant Tuberculosis(MDR-TB); a TB strain that is highly resistant to TB drugs and the needs of the poor and vulnerable populations; (iii) contributing to health-system strengthening based on primary health care; (iv) engaging all care providers; (v) empowering people with TB, and communities through partnership; and (vi) enabling and promoting research.

The *Mycobacterium tuberculosis* genome was first completely sequenced just over a decade

ago in 1998 Cole et al. (1998), long after Robert Koch, a bacteriologist, isolated the infectious agent (tubercle bacilli) in 1882. This feat in the medical field later earned him a Nobel prize in medicine in 1905 Koch (1905).

TB is a very difficult disease to manage in the sense that, clinically it presents with symptoms that are similar to other diseases such as Malaria, and false negative skin-test results may be triggered by such infections as measles (http://www.australianprescriber.com/magazine/33/1/12/18/). In addition, in MTB culture which is the definitive way of diagnosing TB, sensitivity at 75-80% is greatly affected by the slow division of the MTB bacteria (once every 20 hours), which results in the MTB culture taking up to 4-6 weeks to grow Cox (2004). This is a lot of time to allow the bacteria to establish itself in its host. The fact that TB sufferers need to follow a long and strict drug regimen which may last over six months once diagnosed with TB is also a major blow to the TB control strategy WHO (2009).

Complete sequencing of MTB as with any other organism, provides researchers with the opportunity to analyze the different functions that the complete set of genes in its genome perform Puig et al. (2001). In order to understand the complex biological systems that organisms are, there is a need to analyze their gene regulatory networks in addition to fully understanding the identity, modification and expression levels of encoded proteins in the organism's genome Puig et al. (2001).

One way to better understand the virulence of *M. tuberculosis* is to investigate the molecular interactions that come into play when the pathogen infects its host. This can be achieved by elucidating the protein-protein interactions (PPIs) both among the proteins constituting the complete proteome and those that occur between host-pathogen protein pairs for the *M. tuberculosis* pathogen and its host, the human cells.

With a clear understanding of a pathogen's PPI network, it is possible not only to identify essential proteins but also to identify the biochemical pathways that they are involved in Shoemaker and Panchenko (2007a). PPI networks together with metabolic pathways that these networks depict can aid researchers in identifying suitable candidates for drug targets.

This study, through the application of bioinformatics techniques, contributes to the research component of the stop TB strategy by analyzing protein-protein interactions

amongst the proteins that comprise the whole *Mycobacterium tuberculosis* proteome.

With technological advancement in biological sciences for instance in such areas as sequencing technologies Schuster (2008), a lot of biological data is generated at an exponential rate thereby demanding response by experts in the field of computational biology and bioinformatics in the form of developing methods and tools to accurately and efficiently translate these data into meaningful biological knowledge.

This research will describe a number of computational methods for predicting and validating protein-protein interactions that occur within the complete genome of *M. tuberculosis*.

## 1.1 Objectives of the research

This research project aims to predict and study protein-protein interactions in *Mycobacterium tuberculosis* through the following steps:

1) Predict new protein-protein interactions for *M. tuberculosis* through protein domain-domain interactions and orthologous protein-protein interactions in other species, and integrate the results with experimental protein-protein interaction data, subcellular localization data, and functional interaction data from STRING
von Mering et al. (2007), a database that catalogues known and predicted functional interactions of proteins.

2) Secondly, we aim to develop an integrative scoring mechanism that takes into account all the evidence supporting a particular interaction and assign final evidence scores based on both the weight of the evidence(s) and the number of methods that support that interaction. Ultimately we aim to generate an interaction network of all the possible interacting proteins with scores supporting such interactions.

3) The third aim of this study is to find the evolutionary relationship between each pair of interacting proteins. Specifically, this part seeks to answer the question as to whether interacting proteins follow a concerted evolutionary pattern.

4) Lastly, we aim to investigate some of the predicted interactions for biological relevance in light of the evidence that supports them. Here we investigate the shared annotations

among interacting proteins of subcellular localization and biological processes. We also aim to identify significantly overrepresented functions in the set of predicted PPIs.

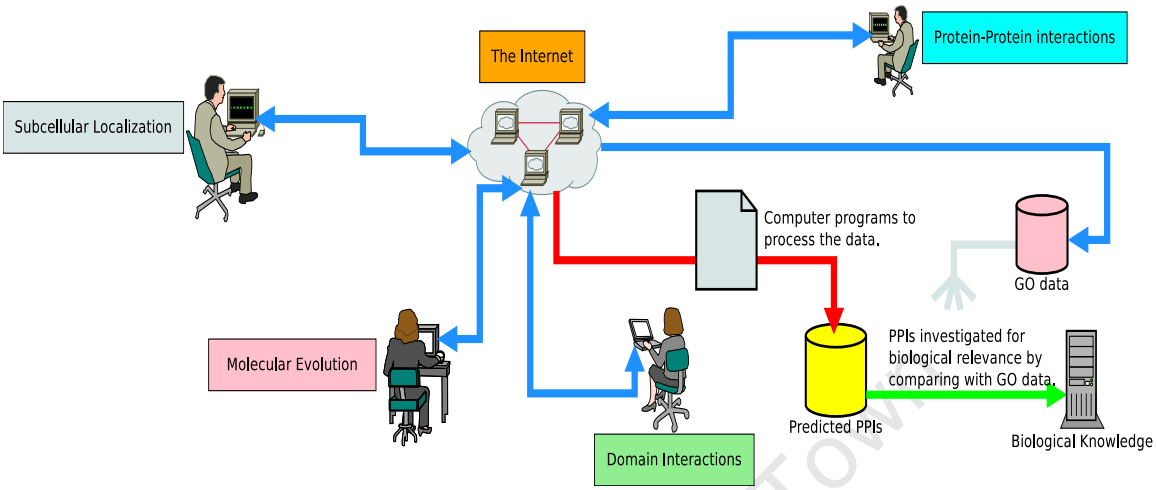


Figure (1.1): A summary of the process involved in predicting protein-protein interactions. The figure above shows in a diagrammatic summary, the process that we use to identify protein-protein interactions (PPIs) and at the end derive biological knowledge from the predicted protein-protein interactions. Experts in different spheres work in different research areas including (sub cellular localization prediction, domain interactions, protein-protein interactions and molecular evolution) and post their findings and tools online. Our strategy involves predicting protein-protein interactions *in silico* from the data retrieved from the research community and identifying all the evidences that support the predicted interactions. The predicted PPIs are scored according to the level of evidence that supports them. Finally, we investigate the predicted PPIs for biological relevance and the findings are stored as biological knowledge.

Figure 1.1 provides a summary of the different contributing data sources for the project and how these are integrated with biological data to predict and study protein-protein interactions in MTB. The different research areas or data sources are considered to be independent of one another.

## 1.2 Road map

Chapter 2 covers the background to this research work by looking at the various protein-protein interaction prediction methods. The chapter starts with a brief introduction to protein sequences while focusing on their properties that are pertinent to this research in a more specific manner including: 1) protein domains; 2) protein sub cellular localization and 3) protein molecular evolution. We take a critical look at the methods that have been used before to predict protein-protein interaction with emphasis on the methods employed by the databases that we use in this research.

Secondly, we take an in-depth look at molecular evolution of protein sequences with the aim of deciphering the relationship between a protein's evolution compared to those of its interaction partners.

Chapter 3 describes the methods that we used to predict protein-protein interactions in the *M. tuberculosis* genome. We integrate data from gene ontology (GO) in a bid to find GO terms that are highly enriched for every interacting protein pair. We also use data from the cellular component ontology of GO to enrich the subcellular information for the predicted interacting proteins. In this chapter we also describe how we study the molecular evolution of the predicted interacting protein pairs and outline the methods for calculation of dN/dS and and codon volatility values for each predicted interaction. We conclude the chapter by describing a scoring scheme that takes into account all the evidences used for each predicted interacting protein pair, as well as the methods used for biological interpretation of the results.

Chapter 4 presents the results that we obtain from the prediction algorithms that we develop in Chapter 3. One of the ultimate objectives of this research project is to come up with an interaction network, complete with scores supporting each interaction pair. We outline the interaction network in form of PPI pairs with complete scores and evidence that supports the interaction. We also integrate gene ontology data here in a bid to understand whether interacting proteins are enriched for particular functions.

Chapter 5 discusses the results of this research in-depth. We compare the results of protein-protein interactions inferred from different methods, and more importantly, how they fit in the current knowledge base of TB. The chapter takes a critical look at some of the

predicted interactions with the aim of finding out whether some of the novel interactions have biologically meaningful interpretations with regards to feasibility of such interactions as well as the strength of the evidence that support them.

In Chapter 6 covers the conclusions that we draw from this research in light of the objectives set out for this study.

# Chapter 2

# Background



## 2.1 A brief introduction to proteins

This dissertation focuses heavily on proteins, more specifically on protein-protein interaction principles and methods for predicting these interactions. This section briefly introduces protein sequences and some of their properties as well as their elements that the dissertation handles at length.

Proteins are one of the four building blocks of biological life forms, the others being carbohydrates (sugars), lipids (fats) and nucleic acids (DNA and RNA).

Proteins perform various functions in an organism, including forming parts of structural elements of the cell, some proteins work as enzymes that catalyze various biochemical reactions in the body, others function as antibodies to help an organism's immune system fight off infections, and other proteins function as hormones that send signals throughout an organism's system.

In a little more detail, proteins are biological macromolecules formed by a linear chain of amino acid residues linked by peptide bonds. A protein's primary structure, also referred to as its sequence, is the linear order of amino acid residues from the amino- (or N-) terminus to the carboxy- (or C-) terminus of the protein chain. It is also important to note that protein sequences are encoded by DNA. The DNA is transcribed to RNA by RNA polymerases, which is then translated to protein via ribosomes. The above two processes;

transcription and translation, sometimes referred to as the Central Dogma of Molecular Biology, are key to all known cellular life on this planet.

Codons (trinucleotides) in a coding region of DNA encode twenty amino acids that form building blocks in most species. There are a set of 61 codons that are mapped to 20 amino acids and 3 stop-codons, so called as they signal the stopping of translation. These 61 codons are collectively referred to as the genetic code. While the DNA and RNA hold the information of a cell, proteins are important to organisms, functioning as structural components, signaling molecules, enzymes, transcription factors and receptors, among other functions.

A protein sequence is computationally represented as a string of characters each representing a one-letter abbreviation of the corresponding amino acid at that position.



### 2.1.1 Protein domains

Domains as a concept in protein studies was first proposed in Wetlaufer (1973). Wetlaufer described domains as those regions within proteins (always between 40 - 150 amino acids long) which can fold autonomously to form stable protein structures.

Various researchers have in the past described domains in different ways including structurally Richardson (1981), by function and evolution Bork (1991), and also by considering how they fold Wetlaufer (1973). These descriptions are all valid, as when taken separately they support each other and even overlap in many ways.

Consequently, a protein domain can therefore be described as an evolutionary conserved region on the protein sequence that is functionally and structurally independent of the entire protein Ng et al. (2003b). Sikder and Zomaya (2008) defined a domain as the fundamental unit of protein structure, function, folding, evolution, and design.

Protein domains should however be distinguished from protein motifs, which even though they are evolutionary conserved, at ten to twenty amino acid residues long they are much shorter than protein domains which are normally 40 to 150 amino acid residues long with an average length of 100 amino acid residues. Protein motifs are also known to be associated with distinct structural sites that perform a particular function.

Protein domains are important in the sense that it is the domain structure of a protein that determines a protein's function, the biological pathways in which it is involved, and the molecules that it interacts with Sikder and Zomaya (2008). Sikder and Zomaya (2008), also showed that structurally similar domains often occur in different proteins in spite of the protein sequences showing no noticeable similarity. The knowledge of a protein's domain structure therefore provides vital leads as to the functions of a protein and its possible interaction partners, which is key in proteomic studies Sikder and Zomaya (2008).

Proteins can thus be viewed to be composed of a finite set of domains which are joined together in diverse combinations. Due to their ability to exist independently of the entire protein, domains are important in drug discovery studies as their contribution to the function of the entire protein can be studied and analyzed in isolation from the rest of the protein Sikder and Zomaya (2008).



### 2.1.2 Protein subcellular localization prediction

Subcellular localization of a protein can be viewed as the location of a protein within one or more (for transmembrane proteins) of the membrane bound regions within a cell or in some cases outside the cell, in the case of exported proteins. Protein subcellular localization prediction involves experimental or computational prediction of where a protein resides in the cell.

It is important to study a protein's subcellular localization as it not only gives insight into what a protein's function might be Emanuelsson et al. (2007), but also gives vital leads as to how the cell as a whole is organized more so, when the localization of proteins within different compartments of a cell are studied Scott et al. (2005).

Several laboratory techniques such as immunofluorescence and electron microscopy Kumar et al. (2000), fluorescent-protein tagging Kenri et al. (2004), and the Western/SDS-PAGE analysis Hancock and Nikaido (1978), have been used to identify protein subcellular localization. These methods provide relatively high-quality localization information however, they are limited by the fact that they are applicable to single proteins, or at best small sets of proteins Rey et al. (2005). The small amount of protein data handled by these methods also impedes their popularity due to the resulting high cost in experimental design and a lot of time that is spent in cases where subcellular localization information is needed for a

large volume of proteins.

### Methods for identifying subcellular localization

Over the years, several methods have been developed to determine the subcellular localization of proteins using high-throughput experiments. Burns et. al, made one of the first attempts by randomly inserting the *lacZ* reporter gene into the yeast genome, and were able to determine the subcellular localization of a total of 245 yeast proteins Burns et al. (1994). A similar method to the one above, explained in Chalfie et al. (1994) uses green fluorescent protein or simian virus V5 epitope to tag cDNAs and localize the resulting fusion proteins through fluorescence screening of the transfected yeast cells.

Even though the methods mentioned above allow for the study of protein subcellular localization *in vivo*, the presence of fusion/tagging protein may interfere with sequence or structural signals necessary to direct the protein of interest to its proper compartment Emanuelsson et al. (2007). Another approach would be to homogenize and fractionate (through centrifugation) the cell and use mass spectrometry Shevchenko et al. (1996), in order to identify the proteins in the various fractions. A third approach was described in Ramos-Vara (2005), where the authors demonstrated that protein-specific antibodies can be designed and used through immunohistochemistry, to map tissue specificity and subcellular localization of human proteins. High-throughput experimental techniques of predicting protein subcellular localization inevitably produce some false positive assignments (as well as false negatives), quality of protein subcellular localization is much improved when experimental methods are used in conjunction with computational tools that score the likelihood that a protein belongs to a given compartment Emanuelsson et al. (2007). Such scores can be used to improve the quality of high-throughput data, and also subsequently used as starting points for identification of compartment-specific protein complexes or networks Emanuelsson et al. (2007).

In Rey et al. (2005), the authors observed that some computational subcellular localization prediction methods exceed their high-throughput experiment counterparts in prediction of protein subcellular localization. This observation prompts in-depth analysis of computational protein subcellular localization methods and incorporating results in scientific studies. In this research we use the PSORTb 3.0 Yu et al. (2010), tool to predict protein

subcellular localization information for MTB in addition to text mining UniProt files for protein subcellular localization information.

We integrate subcellular localization data to putative protein-protein interaction data that we generate using the methods described in the method section in order to find support for predicted protein-protein interactions.

## 2.2 Protein-protein interactions

A protein-protein interaction (PPI) occurs when two proteins bind together, in most cases to carry out a biological function. Proteins interact to mediate many of the cellular processes in a living organism.

Proteins catalyze reactions, they transport nutrients, form building blocks of viral capsids, traverse the membranes to yield regulated channels, and transmit the information from DNA to RNA Keskin et al. (2008). Some of the most important functions of proteins are as vehicles of the immune response of a host organism and facilitating viral entry into the host organism cells.

The ultimate goal of many protein-protein interaction studies is to be able to assign functions to proteins that interact in a network. Recognition of the fact that proteins are involved in almost all cellular processes has led to focused attempts at predicting their functions from their sequences, or if available, their structures, for example in Young (1998), Mika and Rost (2006), Zhu et al. (2000), Skrabanek et al. (2008). Besides, Valencia and Pazos observed that properties of many complex systems are more closely determined by their interactions rather than by the characteristics of their individual components Valencia and Pazos (2002).

In many cellular processes, proteins recognize and bind to specific targets in a highly regular manner Shoemaker and Panchenko (2007b). The specificity of interactions in these cases is determined by the structural and physico-chemical properties of the two interacting proteins. This as a result, qualifies the need for conservation, to a certain degree of interaction patterns between similar proteins and domains Valdar and Thornton (2001). As evidence of functional relationship between similar proteins, close homologs have been

observed to almost always interact in the same way. Consequently, protein-protein interactions place certain evolutionary constraints on protein sequence and structural divergence in order to maintain these interactions Shoemaker and Panchenko (2007b).

Since the majority of protein functions in a living cell are mediated by protein-protein interactions, if the function of at least one of the components with which the protein interacts is identified, then functional and pathway assignment for the remaining protein in question is facilitated. A practical way to predict a protein function is thus by identifying its binding partners, otherwise known as guilt by association Oliver (2000).

It is therefore important to identify and characterize protein-protein interaction networks in order to understand, on a molecular level the mechanisms of the biological processes that occur in a living cell Shoemaker and Panchenko (2007b). When dealing with a pathogen, knowledge about its protein-protein interaction network becomes key, especially when identifying possible drug targets when designing drugs that inhibit that pathogen. Through a protein-protein interaction network, we can map cellular pathways and their intricate cross-connectivity Keskin et al. (2008). The knowledge of how cellular pathways interact is important in inferring their dynamic regulation which is very important in drug discovery.

One of the challenges that face functional annotation of genomes is the slow pace at which protein-protein interaction are identified, which according to von Mering et al. (2007), is not abreast with the pace of genome sequencing. This calls for faster and more reliable protein-protein interaction detection methods which, in addition, can help in validating already identified interactions. Researchers have resorted to different methods to tackle this challenge by, developing better experimental approaches and computational methods that have not only proved to be faster but are also gaining popularity due to their relatively low cost in comparison with experimental methods.

Computational prediction of protein-protein interactions consists of two main areas (i) the mapping of protein-protein interactions, i.e., determining whether the two proteins are likely to interact and (ii) understanding the mechanism of protein-protein interactions and the identification of residues in proteins which are involved in those interactions Skrabanek et al. (2008). The first computational analysis of protein-protein interactions used the structural context of proteins to analyze known protein-protein interaction interfaces in order to determine physical rules determining protein-protein interaction specificity Skrabanek et al. (2008). The physical structure that two proteins achieve determines largely the

ease with which they can bind to each other and hence form a protein-protein interaction.

In Chothia and Janin (1975), the authors first attempted to describe the characteristics of protein interaction sites. They used data of three protein complexes to suggest that residues that form the interaction interface are closely packed and tend to be hydrophobic and that complementarity may be an important factor in predicting which proteins can interact.

Later studies, which involved the use of even larger datasets developed and extended Chothia and Janin's work to try to identify characteristics of the interaction sites that are sufficiently different from the rest of the protein to be identifiable, and thus predictive. Analysis of the hydrophobicity distribution of amino acids can be used to predict interaction since interacting regions tend to be the most hydrophobic clusters on the surface of the protein Young et al. (1994), Mueller and Feigon (2002).

Structural methods to predict PPIs in MTB are limited by the fact that not all know MTB transcripts have their structures solved. According to ModBase database data (September, 2011), about 70% of MTB transcripts have their structures modelled Pieper et al. (2011).

Protein interactions fall into different types depending on their strength (permanent and transient), specificity (specific and non-specific), the location of interacting partners within one or two polypeptide chains, similarity between interacting subunits (homo- and hetero-oligomers) Shoemaker and Panchenko (2007b).

Several methods including Valdar and Thornton (2001), Rigaut et al. (1999), Young (1998), have been applied previously in identifying protein-protein interactions. The methods that have traditionally been used to infer protein-protein interactions have involved a top-down, hypothesis driven approach in which scientists design focused experiments to test their hypothesized interactions Ng et al. (2003b). These approaches are time consuming and expensive to carry out in terms of machines needed for experimental setup and personnel required to design and run these experiments.

High-throughput protein-protein interaction detection methods such as two-hybrid systems Uetz et al. (2000) and protein chips Zhu et al. (2000) have been developed to detect interactions on a larger and more rapid scale. These high-throughput methods have however compromised on the quality of interaction data thus generated resulting in high error rates

Braun et al. (2009). Researchers are then faced with both computational and experimental challenges of developing methods that can not only reliably and efficiently characterize detected interactions but also validate these interactions Keskin et al. (2008).

In this study, we explore an integrative approach to inferring putative protein-protein interactions from different data sources such as domain-domain interaction data, subcellular localization data, and functional interaction data from publicly available protein-protein interaction databases including, Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) von Mering et al. (2007), Database of Interacting Proteins (DIP) Xenarios et al. (2000), and IntAct Aranda et al. (2010).

### 2.2.1 Motivation for integrating data to infer protein-protein interactions

When doing comparative studies of large-scale protein-protein interactions predicted by disparate methods, von Mering and colleagues found that each method that was used in predicting protein-protein interactions produced a unique distribution with respect to functional categories of the interacting proteins von Mering et al. (2002). They for example, found out that datasets based on purified complexes predict relatively few interactions for proteins involved in transport and sensing. A possible explanation for this occurrence would be that transport proteins are enriched in transmembrane proteins which are more difficult to purify. In the same respect, interactions detected by yeast two-hybrid technology largely fail to cover certain categories; for example, proteins involved in translation are found comparatively less often than by other methods.

The observations made above suggest that different methods have specific strengths and weaknesses with respect to predicting protein-protein interactions of proteins belonging to different functional categories. It is therefore important to integrate different methods. Integrating different methods to predict protein-protein interactions not only produces additional validation layers of the plausibility of a predicted PPI being a true interaction, but also counters the false negative predictions that result from the weaknesses of different methods. This consequently, leads to a more comprehensive coverage of protein-protein interactions.

### 2.2.2 Experimental methods for measuring protein-protein interactions

Many methods have been used to measure protein-protein interactions some of which are briefly discussed in this chapter. However, the data obtained by these methods are partial, hence many interacting proteins are yet to be identified according to Ben-Hur and Noble (2005). It is estimated that for a well studied organism like yeast, only about half of the complete interactome have been discovered Ben-Hur and Noble (2005). High-throughput experimental methods as well as computational methods have been used to measure PPIs. However, there have been a few cases of overlap, even for studies using the same method. In fact in a study by Ben-Hur and Noble (2005), an assay sensitivity of only up to 25% failed to detect random reference set (RRS) for the assays in the yeast proteins. This observation suggest that there are still many interactions that are not yet accessible to high-throughput experimental methods Ben-Hur and Noble (2005), hence there is an urgent need to develop better computational methods in order to access these interactions. The relatively less expensive computational methods can then be verified by the relatively expensive and labor-intensive experimental methods. In this study, we briefly explore a background on various methods that are currently used to measure protein-protein interactions–specifically analyzing their merits and flaws.



#### Yeast two–hybrid assay (Y2H)

Y2H is based on the fact that many eukaryotic transcription activators have at least two distinct domains, one that directs binding to a promoter DNA sequence, the Binding Domain (BD), and another that activates transcription, the Activation Domain (AD) as shown in Figure 2.1. When the BD and AD domains are split, transcription is inactivated Shoemaker and Panchenko (2007b).

In this method pairs of proteins to be tested for interaction are expressed as fusion proteins (hybrids) in yeast. One protein is fused to a DNA-binding domain, while the other one is fused to a transcriptional activator domain. Any interaction between them is detected by the formation of a functional transcription factor. Below is a diagrammatic representation of the process of Yeast two–hybrid assay.

The diagram is adapted from http://www.onesci.com/w/wix/index.php?title=Daily_Method&oldid=4468 and Oliver (2000).

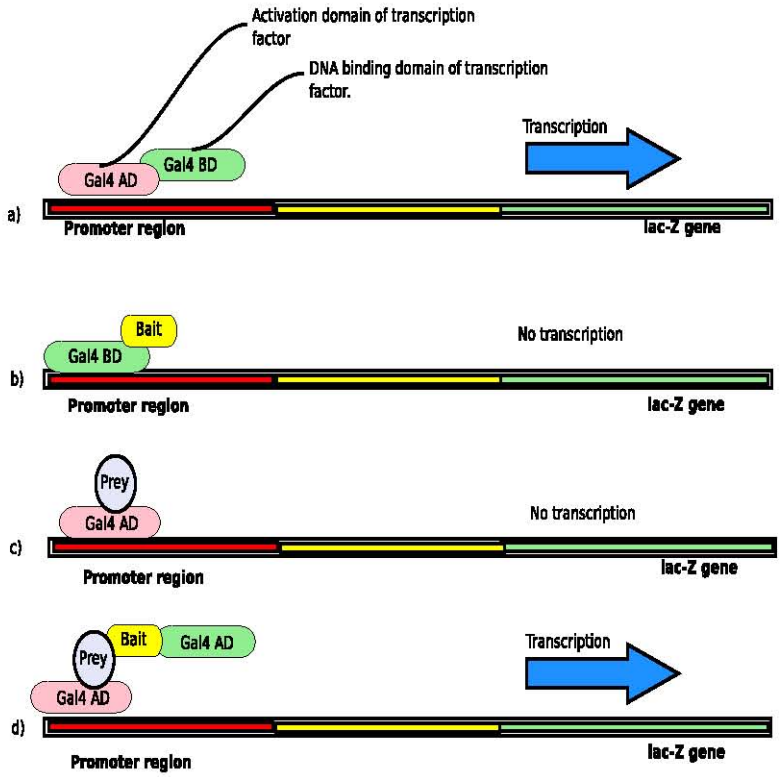


Figure 2.1: Illustration of Yeast Two Hybrid Experiment using Gal4 as a transcription factor. a) The gene that activates transcription of the lac-Z gene occurs as two distinct transcription factor domains namely:Activation domain,(Gal4 AD) and DNA binding domain, (Gal4 BD) of the transcription factor. b) and c) show that no transcription is activated if either of the domains is absent. In d), transcription of the *lac-Z* is activated as a result of the interaction of the 'Bait' and 'Prey' proteins resulting in the formation of a protein complex involving both AD and BD.

*Benefits:* Yeast two–hybrid allows for detection of transient and unstable interactions. Secondly, it is independent of endogenous protein expression, and it has a fine resolution enabling interaction mapping within proteins.

One major disadvantage of the Y2H method is that it is highly prone to false positive

interactions Braun et al. (2009).

#### Purification of protein complexes

In this approach, individual proteins are tagged and used as baits or 'hooks' to biochemically purify protein complexes Meur and Gentleman (2008). The complexes that are formed by the tag and the prey proteins are then purified and the components identified by mass spectrometry Ho et al. (2002), Meur and Gentleman (2008), Rigaut et al. (1999).

The merits of this method are that several proteins in a complex can be tagged, hence it allows for checking internal consistency plus it also draws from the benefits of *in vivo* techniques, as it detects interactions in physiological settings.

The demerits are that it might mix complexes that are not present under given physiological conditions, the process of tagging the proteins may also interfere with the complex formation, and loosely associated complexes may be washed off during purification Meur and Gentleman (2008).



### 2.2.3 Computational methods for predicting protein-protein interactions

This section briefly explains *in silico* methods that have been used to infer PPIs. Further details on the implementation of these methods can be retrieved from the references provided for each of the methods.

#### Gene fusion

The gene fusion strategy posits that interactions between proteins can be deduced from the presence in different genomes of the same protein domains, which can either form part of a single polypeptide chain (multi-domain protein) or act as independent proteins (single domains) Valencia and Pazos (2002). A study by Enright et al. (1999) and Marcotte et al. (1999) suggest that metabolic efficiency in the form of reduced regulation load of multiple interacting gene products is the driving force behind gene fusion events. Therefore, gene

fusion events provide an elegant way to computationally detect functional and physical interactions between proteins.


Figure 2.2: Illustration of gene fusion concept. In the figure protein A and B occur separately in organism X. Whereas in organism Y, they are fused together into a single protein. This observation suggest that protein A and protein B are functionally related therefore interact.

Recursive sequence search and multiple sequence alignment methods are employed to detect domain fusion events Enright et al. (1999), Marcotte et al. (1999). Gene fusion events have been shown to be common in metabolic proteins according to Tsoka and Ouzounis (2000).

Gene fusion is, however, currently restricted only to shared domains in distinct proteins. The limitation with this approach is that the extent of domain overlap between proteins is a phenomenon that is still not very well understood especially in prokaryotes Sprinzak and Margalit (2001). Gene fusion event is illustrated in 2.2 in which the two proteins A and B occur separately in organism X, but are fused into protein C in organism Y.



### Similar phylogenetic profiles

This method is based on the pattern of the presence or absence of a given gene (or set of genes) in a set of genomes Valencia and Pazos (2002), Pazos et al. (2008). In this approach, the similarity of phylogenetic profiles among the genomes is interpreted as being indicative of functional linkage, as the similarity suggests that the genes need to be simultaneously present in order to function together.

The first limitation of the phylogenetic profile method is that it can only be applied to complete genomes, as it is only when you consider the genome in its entirety that can you rule out the presence of a given gene.

Secondly, this method cannot be used with the essential proteins common to most organisms as it may lead to spurious prediction of interactions where in actuality none exists.

### Conservation of gene neighborhood

Bacterial genomes are known to be organized into regions that tend to code for functionally related proteins such as operons Valencia and Pazos (2002). Studies including that of Dandekar et al. (1998), have shown that conservation of gene neighborhood across species can be indicative of physical interaction of proteins that are encoded by the conserved gene pairs. The conservation of gene neighborhood approach has been used to predict functional relationships between adjacent genes in various bacterial genomes Bhardwaj and Lu (2005), Dandekar et al. (1998).

This method generally has the limitation that it is only applicable to bacteria where the order of the genome is relevant Valencia and Pazos (2002). In this study however, this limitation would not apply as we are dealing with a bacterial genome.

### Correlated gene expression

Genes encoding interacting proteins depict strongly correlated expression levels over different experimental conditions in *Saccharomyces cerevisiae* Eisen et al. (1998), Ge et al. (2001), Fraser et al. (2004). A possible explanation for this observation would be that interacting proteins need to be present in the cell in similar amounts at the same time to properly form stoichiometric complexes and execute their function.

### Text Mining

Text mining as a method of identifying PPIs involves extracting PPI data from abstracts or full texts from literature databases such as PubMed, (http://www.ncbi.nlm.nih.gov/pubmed). The process of retrieving the PPI data can either be manually conducted by expertly trained database curators or automated through computer algorithms. Manually curated protein-protein interaction databases include BIND, DIP, HPRD and MINT. An example of a database that uses computational text mining algorithms to retrieve PPI

data is the Online Predicted Human Interaction Database (OPHID), Brown and Jurisica (2005).

Challenges facing text mining as a method of inferring PPIs include the rapid rate at which interactions are reported in PubMed.

### Similarity of phylogenetic trees (mirror trees)

Studies have shown that interacting proteins exhibit correlated evolution Pagès et al. (1997), Fryxell (1996). Pages et al. Pagès et al. (1997), and Fryxell Fryxell (1996), respectively showed that co-hexins co-evolve with dockerins and insulin co-evolves with insulin receptors.

According to Fryxell (1996), gene duplications and functionally related gene families often show similarities in divergence dates, functional specificities, and phylogenetic tree topologies. As an explanation for this observation, Fryxell adds that these correlations suggest that the family trees of functionally related gene families co-evolved because functionally complementary gene duplication and divergence events tended to be retained by natural selection.

Tandem genetic duplications in bacteria and bacteriophages occur spontaneously at a frequency of $10^{-3}$–$10^{-5}$ per locus per generation, and can be of unlimited size. Genetic studies of insecticide resistance on bacteria and bacteriophages have been observed to occur at the same frequency as tandem genetic duplications Fryxell (1996). Fryxell hypothesized that duplicated genes would tend to be lost (or mutated into pseudogenes) unless stabilized by natural selection, which would require some useful functional complementarity with their interacting partners. Fryxell's investigation found for instance that polypeptide growth factors and their receptors co-evolve when they investigated the phylogenetic trees of insulin and insulin receptors Fryxell (1996). The similarity between the trees was found to be higher than expected just from general divergence. The similarity between the phylogenetic trees of interacting proteins was interpreted as an indication of their coordinated evolution and a direct consequence of the similar evolutionary pressure applied to all the members of a given cellular complex.

In cases of phylogenetic tree analysis like the process followed by Fryxell above, correspond-

ing phylogenetic trees of interacting proteins have been shown to display greater similarity (symmetry) than noninteracting proteins would be expected to show Valencia and Pazos (2002). Pazos and Valencia extended the mirror tree procedure to cover large sets of interacting proteins and domains, for which they found out that the value of the correlation between the distance matrices of pairs of interacting proteins was a good indicator of the probability of the two proteins interacting Pazos and Valencia (2001).

According to Pazos and Valencia (2001), an extreme case of co-evolution would be a case where both interacting proteins are simultaneously lost in the same species in the course of evolution. This concerted loss, they said, can be explained by the need by one of the proteins for the other to be present so that it can accomplish the particular function in question. One such example is His5 (His synthesis) and Trp (Trp sythesis).

One main limitation of the mirror tree method however, is the need to obtain good quality, complete multiple sequence alignments MSA for the two proteins. The MSAs should include sequences from different species for the two proteins under consideration.



**In silico two-hybrid method (i2h)**

In silico two-hybrid strategy is to quantify the degree of co-variation between pairs of protein residues (correlated mutations). Interacting proteins have previously been shown to evolve in a correlated manner Fryxell (1996). Conservation and mutation patterns observed in interacting proteins are evidence of functional and structural constraints plus mutational drift Göbel et al. (1994). It is therefore tenable that correlated mutations observed in multiple sequence alignments in a sequence family could be indicative of probable physical contact in three dimensions Göbel et al. (1994).

Valencia and Pazos (2002), suggest that concerted mutations function to be compensatory mutations that stabilize the mutations in one protein with changes in the other. Correlated mutations have in some proteins been shown to select the suitable structural conformation based on the signals at the interaction interfaces Valencia and Pazos (2002), Pazos et al. (1997). The relationship between correlated residues and interacting surfaces has been used in predicting the possibility of interaction between proteins with the analysis based on the differential accumulation of correlated mutations between interacting proteins and within the individual proteins.

A schematic diagram outlining the process of *in silico* two-hybrid is explained in Figure 2.3. The diagram is adapted from Valencia and Pazos (2002).

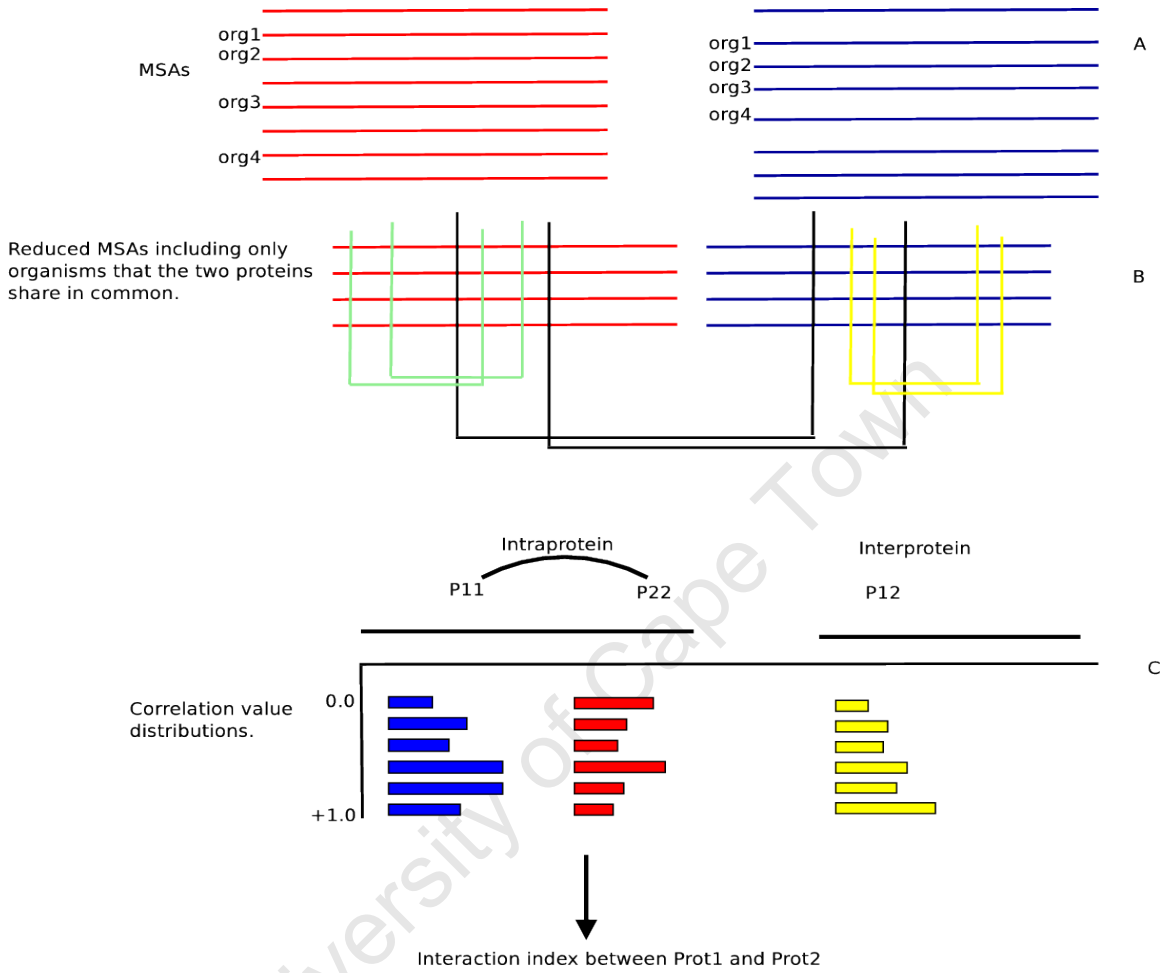




Figure 2.3: Schematic representation of the i2h method. **A**: Family MSA is defined for two protein sequences, 1 and 2. Included in MSA are corresponding sequences from different species represented here by *org1* through to *org4*. **B**: a virtual alignment is constructed, conctenating the sequences of the probable orthologous sequences of the two proteins. The pairs are divided into three sets: two for the intraprotein pairs (P11 and P22; pairs of positions within Prot 1 and within Prot 2) and one for the interprotein pairs (P12 **C**: The distributions of correlation values are recorded for these three sets. Afterwhich, 'interaction index' is calculated by comparing the distribution of interprotein correlations with the two distributions of intraprotein correlations–the details are explained in Valencia and Pazos (2002).

One limitation of the i2h method, as is with most evolutionary studies is the need for complete multiple sequence alignments (MSAs) with a good coverage of species common to the two proteins under study.

### 2.2.4 Challenges in computational prediction of interactions between proteins

Evaluation of PPIs derived computationally suffers from such challenges as limited availability of collections of interacting proteins as well as inaccurate understanding of proteins that do not interact Valencia and Pazos (2002). Clearly, with no rich set of proteins that do not interact (true negatives) *in vivo* there is no possible way of ruling out some interactions, however how unlikely they may seem, more so when dealing with an organism in which majority of the proteins have not been characterized, like in the case of *Mycobacterium tuberculosis*. For example, Yu et al. (2006) compared high-throughput and low-throughput experimental data by examining PPIs of 56 *Saccharomyces cerevisiae* proteins, for which there were complete matrices of experimental results – in other words, proteins for which it is clearly known whether they interact or not. They found out that the experimental results (both high-throughput and low-throughput) agreed on 1033 of all the 1596 all possible interactions, including self interactions. This accounts for about 65% of all the possible interactions. Of the remaining 563 cases which the different experimental results did not agree on, 92.5 % were false negatives while the remaining 7.5 % were false positives. These results showed that high-throughput methods are prone to false negative predictions.

Computational prediction of PPIs is also met with the spatio-temporal challenge of effective interaction. Basically, two proteins even though structurally shown to have the possibility of interacting may not interact *in vivo* due to the requirement that they be present in the same place at the same time Skrabanek et al. (2008). This requirement is not met by many computational methods.

Some researchers have approached the computational prediction of PPI challenge by the application of data-mining techniques on the already available vast collections of biological literature from for example PubMed (`http://www.ncbi.nlm.nih.gov/pubmed`) Blaschke et al. (2002). These systems even though to an extent contribute towards the characterization of PPI, are still faced with such technical challenges as absence of standard protein and gene names as well as complex functional relationships that interacting proteins exhibit Valencia and Pazos (2002).

## 2.3 Co-evolution analysis

Co-evolution at the species level, is defined as the evolution of a species in response to selection imposed by another species Thompson (1989). Recent studies have identified co-evolutionary relationships in different areas including inter-species competition for resources observed between species that share ecological niches, parasite and prey, parasites and hosts; and also symbiotic relationships. The examples for the above mentioned relationships can be found in Moya et al. (2008). Co-evolving species have been observed to share substantial similarities in their evolutionary histories, this is seen for example in the congruence of their phylogenetic trees Pazos and Valencia (2008). Following the similarity observed between phylogenetic trees of co-evolving species, Pazos and Valencia (2008), used the term 'co-evolution' to refer to similarity of evolutionary histories.

### 2.3.1 Background to co-evolutionary analysis

Dependencies between amino acids and proteins have previously been used to unearth the functional relationships between proteins and amino acids Codoner and Fares (2008). One way to find out the dependency between a pair of protein sequences is by finding how they co-evolve with each other.

Several factors account for the evolutionary relationships between amino acids, co-evolution between two proteins can in fact be classified as being functional co-evolution, interaction co-evolution or stochastic co-evolution Codoner and Fares (2008).

A concerted evolution between a pair of proteins has been shown in some cases to indicate a functional relationship between the two proteins as observed in Yeang (2008), Fraser et al. (2004), Fares and McNally (2006).

We define molecular evolution data under two broad subjects based on the methods used to derive the molecular evolution information. The methods that we use to derive molecular evolution data are based on the following two principles: phylogenetic analysis; and codon volatility.

The phylogenetic analysis method works on the widely applied concept of calculating the

ratios of non-synonymous to those synonymous changes (dN/dS) across all of the genes to estimate selection pressure occuring on the genes Pond et al. (2005), Yeang (2008).

The codon volatility method on the other hand uses footprints of non-synonymous substitutions on genes to estimate selective pressures of genes relative to that of the entire genome Plotkin et al. (2004). Codon volatility can also be used to find genes that show significantly more, or less, pressure for amino acid substitutions. These two methods are explained in a little more detail in Chapter 3.

The MSA method basically involves aligning sequences under investigation and finding out from the alignment the sequence mutational dynamics of the aligned sequences.

Estimation of synonymous (dS) and non-synonymous (dN) substitution rates provides an important means of understanding the mechanisms of molecular sequence evolution. A dN/dS ratio significantly greater than one is a convincing indicator of positive selection Yang (1998), a dN/dS ratio equal to one indicates neutral selection and a dN/dS value of less than 1, indicates purifying selection Yang (1998). Mutation and selection (whether positive or negative) have different effects on these substitution rates.



### 2.3.2 Importance of co-evolution

Functionally related amino acid residues are so tightly evolutionarily linked as to preclude dramatic effects on dependent amino acid positions. Due to this interdependence, the selection coefficient against changes in one amino acid site may be highly correlated with the complexity of its intra-molecular interaction networks Codoner and Fares (2008). A good strategy to unearth these relationships is by calculating the relative rates of amino acids at these amino acid sites in a bid to decipher any correlation.

Detecting co-evolving amino acid sites has been regarded as a good strategy for; (i) functional annotations of protein encoded by unknown genes; (ii) revealing possible interactions between amino acids in the same protein; (iii) predicting protein-protein interactions; and (iv) understanding how complex machineries undergo adaptive changes without having a meaningful effect on the organism Codoner and Fares (2008).

### 2.3.3 Co-evolution of proteins

Interacting proteins have been observed to co-evolve as evidenced by the qualitative similarities between their phylogenetic trees (e.g. insulins and their receptors, dockerins/cohexins and vasopressins/ vasioressin receptors) Pazos and Valencia (2008). The use of the similarity of phylogenetic trees method to infer co-evolution is otherwise known as *mirrortree*.

Studies by Goh et al. (2000), have also supported co-evolution between interacting proteins as seen from the high Pearson's correlation coefficient obtained from the sequence similarity matrices of different pairs of interacting protein families (correlation coefficient values that are as high as 0.86 in a 0–1 scale were obtained from such families as NuoE and NuoF subunits of *Escherichia coli* NADH complex).

Similarity of phylogenetic trees as an inference method to co-evolution is however not without demerits. The most profound challenge being that of constructing good phylogenetic trees which accurately depict the evolutionary relationship between the concerned sequences Pazos and Valencia (2008). In addition, the processes of orthologue detection, distance estimation and tree generation are not trivial.

Another case of co-evolution is the similarity of phylogenetic profiles. Extreme cases of co-evolution have been observed to involve the simultaneous loss of two interdependent proteins across different genomes Pazos and Valencia (2008). An example is a case where one of two proteins that work together is lost for some reason, leading to the subsequent loss of the other. A concerted evolution between proteins is explained by the hypothesis that there is reduced evolutionary pressure to maintain the latter since it can not work alone.

As observed above, many methods that seek to understand the evolutionary dynamics of organisms rely heavily on the examination of Multiple Sequence Alignments (MSA's), hence the quality of the evolutionary studies thus inferred largely depend on the quality of MSAs in terms of size (more aligned sequences more information obtained) and the background noise, which should be as limited as possible Codoner and Fares (2008).

# Chapter 3

# Methods



This chapter discusses two methods that we implement in predicting protein-protein interactions, namely: Domain Evidence Algorithm (DEA) and Ortholog Prediction of Interaction Algorithm (OPIA). The two methods are implemented within a library of functions implemented in the Python programming language (http://www.python.org/). The basic idea behind DEA is that domains that have been observed to interact in other organisms, or mediate protein-protein interactions in those organisms, can be predicted to interact in MTB proteins that have them as part of their constituent domains. Thus MTB proteins that contain these interacting domains can be inferred to interact too.

OPIA utilizes the concept of interologs to predict interaction between MTB proteins whose orthologs interact in different organisms. We collected all the interacting pairs of proteins from the IntAct database using the url http://www.ebi.ac.uk/intact/main.xhtml (March, 2010). For each interaction pair we identified orthologs of both of the proteins constituting the pair in MTB. We inferred that these orthologs in MTB also interact.

We identify the evolutionary relationship between the predicted pairs of interacting MTB proteins by: (1) calculating the non-synonymous to synonymous substitution ratio for the predicted interacting proteins; (2) calculating the codon volatility values for all the proteins in the MTB genome and then extrapolating the codon volatility values to the interaction network.

We develop a scoring scheme for calculating the confidence which we attach to the predicted interactions. The scoring scheme takes into account the number of methods that support a

particular interaction, which are added integratively to derive the final score. We describe a scoring scheme that scores the functional similarity between interacting protein pairs.

We ran our set of predicted interactions through a gene set enrichment pipeline with the aim of finding functional categories that are overrepresented. This process aids in deriving biological meaning from the set of predicted protein-protein interactions.

## 3.1 Identifying protein-protein interactions

In this section we describe the algorithms that we use for inferring PPIs. OPIA relies on the concept of conservation of function across related proteins Kotelnikova et al. (2007). Some studies have used the concept of sequence similarity to infer PPIs, examples of which include domain methods Sprinzak and Margalit (2001), gene fusion Marcotte et al. (1999) and pairwise kernels Ben-Hur and Noble (2005). DEA borrows from the knowledge that PPIs are domain-domain interactions occurring among the constituent domains of the proteins involved in the interaction. We also add data for functional interaction of proteins from the STRING database von Mering et al. (2003a; 2007).



### 3.1.1 Domain-domain interaction

Protein domains are the structural and functional building blocks of proteins Finn et al. (2010). Protein-protein interactions, the main focus of this research, involve physical interactions between the proteins' domains. In most cases, protein binding is characterized by specific interactions of evolutionary conserved domains Weiner and Bornberg-Bauer (2006). In effect, it is more accurate to say that Protein A interacts via Domain $\alpha$ with Domain $\beta$ in Protein B, where $\alpha$ and $\beta$ are domains in protein A and protein B respectively. As a result, protein-protein interactions, at the domain level can be looked at as just domain-domain interactions of the domain pool derived from the interacting proteins. In addition, important information on the cellular function of protein interactions and complexes can often be obtained from the known functions of the interacting protein domains Albrecht et al. (2005).

Domain-domain interactions are graphically explained in figure 3.1

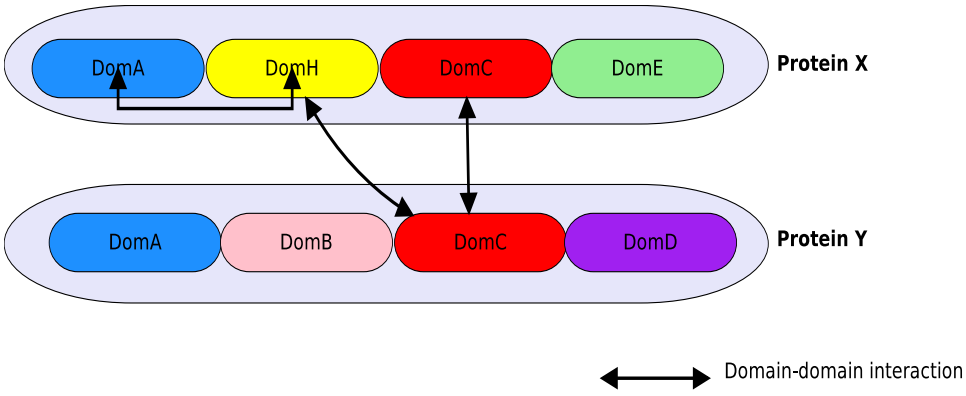


Figure 3.1: Domain-domain interaction between and within proteins. (a)DomA of Protein X interacts with DomH of the same protein (b) DomH of Protein X interacts with DomC of Protein Y (c) DomC of Protein X interacts with DomC which also occurs in protein Y. Domain-domain interactions depicted in (a) represent domain-domain interactions that form multi-domain proteins whereas (b) and (c) above can be used to predict protein-protein interaction between Protein X and Protein Y.



The majority of proteins (two-thirds in prokaryotes and four-fifths in eukaryotes) are multi-domain proteins Raghavachari et al. (2008). Either a transient or a stable interaction between two proteins would involve the physical binding of two or more domains. Thus, understanding domain-domain interactions provide a better path towards understanding precise atomic details of protein-protein interactions.

The conventional data source for deriving domain-domain interactions is from pair-wise protein-protein interactions Ng et al. (2003b). This method has been used in previous work by Wojcik and Schächter (2001), and Deng et al. (2002). In Wojcik and Schächter (2001), the authors showed that the use of domain-domain interactions for *in silico* prediction of protein-protein interactions performs better than the use of full-length proteins.

Other methods involving the use of protein domain interactions to explore protein-protein interactions have also been explored. For instance Gomez et al. (2003), explored the use of domain interaction with network topology to predict protein-protein interactions statistically. Deng et al. (2002) devised a maximum likelihood method to infer domain-domain interactions that they further used to predict protein-protein interactions.

Looking further afield, Ng et al. (2003b) modified the gene fusion method (also known as 'Rosetta Stone') popularised by Enright et al. (1999), and Marcotte et al. (1999), to infer

domain-domain interactions from sequences in different species. The basic idea behind gene fusion is that two genes which occur separately in one species, when observed to be fused into one gene in another species may lead to the plausible conclusion that the two genes work together and therefore interact. It has been suggested that the driving force behind gene fusion events is to lower the regulation load of multiple interacting gene products Enright et al. (1999), hence gene fusion events provide an elegant way to computationally detect functional and physical interactions between proteins.

If the two genes interact to form a complex, then they do so through their constituent domains Deng et al. (2002). Therefore, it is valid to say that their domains interact. Ng et al. (2003b), developed a probabilistic measure that calculated the probability that two proteins whose gene products are observed to be fused in one species interact. The probabilistic measure has a direct relationship with the number of constituent domains making up the two interacting proteins. Ng et al. (2003b), modified the idea of gene fusion to infer domain-domain interactions by suggesting that if two genes interact to form a multi-gene complex then they do so through their constituent domains which interact in order to achieve the fusion event.

One way to infer domain-domain interactions is by studying three-dimensional structures that the proteins form in a bid to understand whether the three-dimensional structures thus formed are likely to interact spatially Raghavachari et al. (2008). Databases have been developed to aid researchers in studying domain-domain interactions. Two databases that use a protein's three dimensional structure to infer domain-domain interactions are iPfam Finn et al. (2005) and 3did Stein et al. (2005). Other databases such as DIMA Ng et al. (2003b), DOMINE Raghavachari et al. (2008), and InterDom Pagel et al. (2008), store computationally predicted domain-domain interactions as well as known domain-domain interactions in some cases.

Protein domain-domain interactions however, unlike protein-protein interactions, have no high-throughput results that are currently available Ng et al. (2003b). In this study we use putative domain-domain interactions derived from the InterDom database. InterDom is a database of putative domain- domain interactions that have been inferred by different methods including protein complexes, domain fusions, protein-protein interactions and scientific literature from different data sources.

If two proteins $Pr$ and $Ps$ are known to bind to each other, Ng et al. (2003a), inferred

that domain $Dr_i$ potentially interacts with domain $Ds_j$ with a minimal probability of $\frac{1}{m_r m_s}$ where $m_r$ and $m_s$ are the number of domains in proteins $Pr$ and $Ps$ respectively, and $Dr_i$ and $Ds_j$ are the $i^{th}$ and the $j^{th}$ domains of the proteins $P_r$ and $P_s$ respectively.

$$(Dr_i, Ds_j) = \frac{1}{m_r}.\frac{1}{m_s} \tag{3.1}$$

According to Kundrotas and Alexov (2007), proteins in some cases interact to form multi-protein complexes as opposed to only binary interactions that are detected by high-throughput methods such as yeast-two-hybrid. In a case that proteins interact to form complexes, suppose proteins $P_1, ..., P_N$ interact to form an $N$-protein complex, we can infer that the $i^{th}$ domain of protein $Pr$, $Dr_i$ and the $j^{th}$ domain of protein $Ps$, $Ds_j$ interact with the minimal probability given by the following equation adapted from Ng et al. (2003a).

$$(Dr_i, Ds_j) = \binom{N}{2}^{-1}.\frac{1}{m_r}.\frac{1}{m_s} \tag{3.2}$$

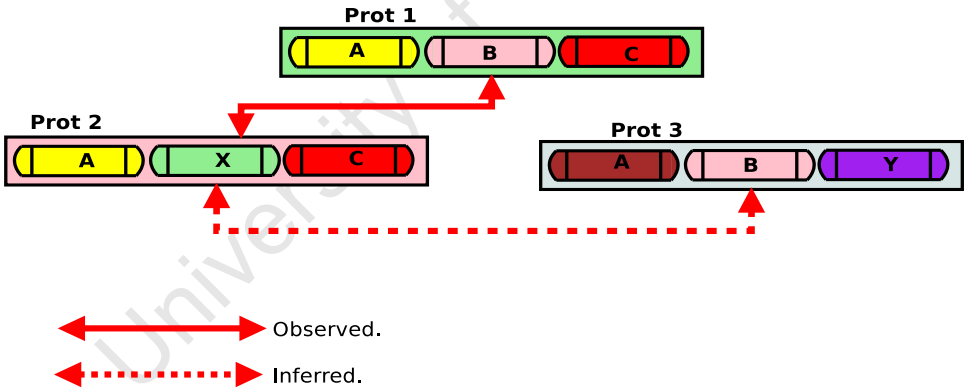


Figure 3.2: Illustration of domain-domain interactions mediating protein-protein interactions. Prot1 interacts with Prot2 is a general way of describing the protein-protein interaction between the two proteins. Specifically, you can say that domain B of Prot1 interacts with domain X of Prot2. Following this train of argument, Prot3 possibly interacts with Prot2 as it has as one of its constituent domains, B, which has been observed to interact with domain X of Prot2.

Figure 3.2 shows graphically how protein domains mediate PPIs.

### Domain Evidence Algorithm (DEA)

We outline the DEA below and graphically in Figure 3.3.

- We first downloaded domain interaction files from the InterDom database at http://interdom.i2r.a-star.edu.sg/.
- The InterDom files can be saved as comma separated files (CSV) where each column stores the following data in the order that the columns appear from left to right.
    1. *domain 1.* First domain.
    2. *domain 2.* Domain (predicted to interact with the first domain).
    3. *gene fusion.* 'yes' if the predicted interaction is supported by gene fusion evidence.
    4. *PDB* Stores the Protein Databank identities of structures where the interaction was observed.
    5. *DIP* Stores DIP interaction identifier (where present) for the predicted interaction.
    6. *BIND* Stores BIND interaction identifier (where present) for the predicted interaction.
    7. *Score* of the interaction as generated by InterDom researchers.
    8. *single int* has the value 'yes' if the predicted interaction is obtained from single domain proteins.
    9. *false positive.* has the value 'yes' if the predicted interaction is a potential false positive according to the thresholds set by the team at InterDom.
- We parsed the InterDom file using a parser program, written in the Python programming language. We retrieved rows that satisfied the following conditions:
    1. Supported by gene fusion evidence.
    2. Supported by a PDB entry.
    3. Is not a potential false positive.

- For the selected domain-domain interactions, we first mapped the PFAM domains Finn et al. (2008) onto InterPro domains Hunter et al. (2009), since we had information on which InterPro domains the *Mycobacterium tuberculosis* proteins that we used in this study had hits on (matched) from the 'InterPro' hit file downloaded from the European Bioinformatics Institute FTP site.
- For each identified interacting InterPro domain pair, we identified the set of all *Mycobacterium tuberculosis* proteins having these domains.
- We then generated pair-wise combinations of all the proteins having taken a protein each from the list of proteins containing *domain 1* and the other part of the pair taken from all the proteins containing *domain 2* while ignoring self interactions. We were not considering interactions that led to formation of multi-chain proteins consisting of a repetition of only one protein.
- We predicted that the protein pairs generated in the previous step potentially interact.

An illustration of predicting domain-domain interactions is shown in Figure 3.4.

**Domain-domain interaction data from the Europen Bioinformatics Institute (EBI)**

The DDIs described in this category have been expertly curated from PPIs for which the curators identified the domains that mediated the PPIs.

1. Domain-domain interaction data on EBI IntAct domain-domain interaction datafile.
2. Parse the file– get interaction pairs where both of the domains in the interacting set also have hits in MTB proteins.
3. Get all the MTB proteins that have the domains from the previous step.
4. Predict the sets of proteins above to interact.
5. Eliminate self interactions.
6. Save the set of non-self interactions in the list of predicted interactions.

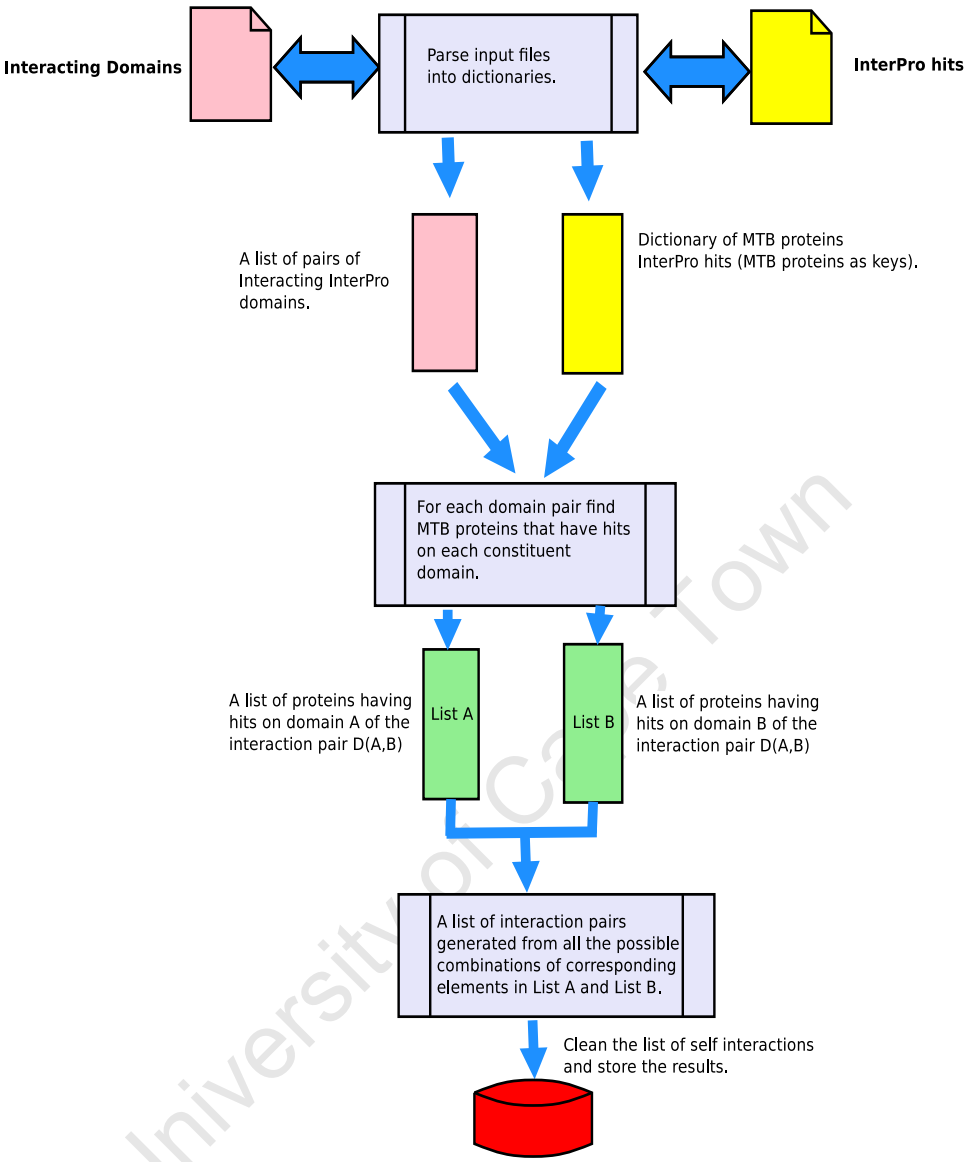




Figure 3.3: Deriving protein-protein interactions from known domain-domain interactions. An illustration of data flow involved in the process of deriving PPIs from known interacting InterPro domains.

### 3.1.2 Interactions inferred from orthologues

IntAct is an open source database of protein-protein interactions that are either expertly curated from the literature or are directly deposited by researchers from protein protein interaction experiments Aranda et al. (2010). Most of the protein-protein interactions in

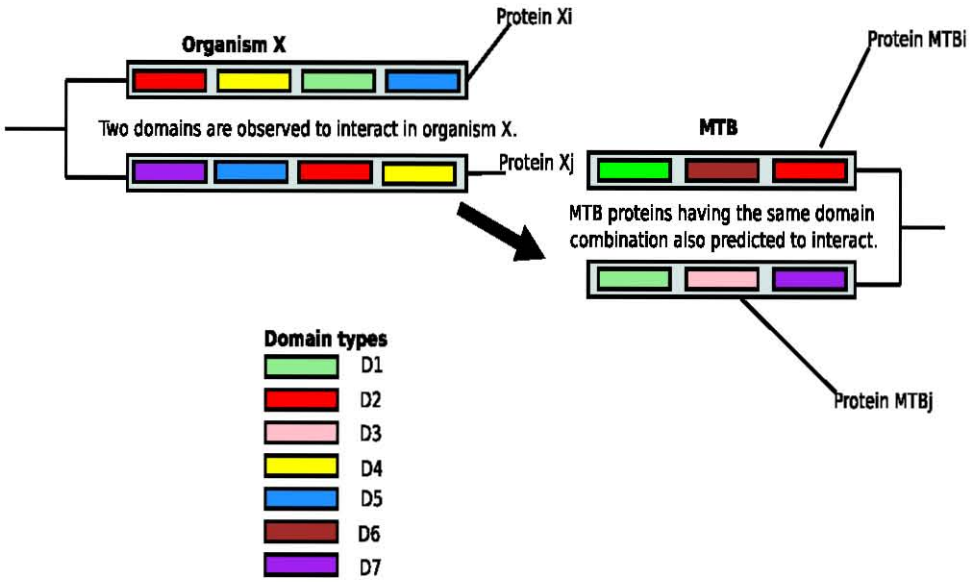


Figure 3.4: Domain-domain interaction transfer. Domains D1 and D7 are observed to interact in organism X, possibly mediating the protein-protein interaction between Protein Xi and Protein Xj. MTB proteins (MTBi and MTBj) that also have domains D2 and D7 in their domain sets are predicted to also interact.

IntAct are observed in yeast *(Saccharomyces cerevisiae)*, fly *(Drosophila Melanogaster)*, and worm *(Caenorhabditis elegans)* which are considered to be model organisms in biological research Fields and Johnston (2005).

We use homology based inferences to infer potential protein-protein interactions in MTB. Homology between two biological entities is simply an evidence of shared common ancestory between them. This could be inferred from significant sequence similarity between the biological entities. Homologous proteins, orthologous proteins in particular, have, in some cases, been shown to conserve protein function. A good example that illustrates the conservation of function and expression phenomenon is described in Rincón-Limas et al. (1999), where in, the authors found that orthologs of the *Drosophila apterous (Ap)* gene in human and mouse effectively regulated the *Ap* target genes in fly, while at the same time producing the same phenotype. The basis of homologous protein sequences sharing functions enables us to infer protein-protein interactions between pairs of *Mycobacterium tuberculosis* proteins from protein-protein interactions derived from its orthologs that interact in other species.

Walhout et al. (2000), developed the concept of interologs as pairs of interacting proteins in one organism whose corresponding orthologs also interact in another organism. Basically, if protein A interacts with protein B in organism X and protein B' interacts with A' in organism Y where A' and B' are orthologs of A and B respectively, then A and A' are interologs and so are B and B'. This idea allows for the transfer of protein-protein interactions for evolutionary related organisms which share sequence similarity and identity among their proteins.

Even though the use of interologs has been in place for some time now, the transfer of protein-protein interactions should be undertaken with caution as there may be cases of false positives in the interaction transfer between different organisms. Yu et al. (2004), for instance, found that only about 16% to 32% of interologs predicted then had experimentally determined interactions correct. This is the result that they achieved when they applied the concept of interologs to *Caenorhabditis elegans* and *Saccharomyces cerevisiae* interacting protein pairs.

For each interacting pair of proteins in the IntAct database, we determined whether there are *Mycobacterium tuberculosis* orthologs for any of the proteins. Identified orthologs for the interacting partners were predicted to interact.

At the time of writing the thesis, there were over 200,000 thousand interaction pairs in the entire IntAct database Aranda et al. (2010). We obtained orthologs of the clinical strain of *Mycobacterium tuberculosis* Fleischmann et al. (2002) from the Intergr8 ortholog file (`http://www.ebi.ac.uk/integr8/EBI-Integr8-HomePage.do`). We inferred MTB orthologs of protein pairs that interact in other organisms to interact in MTB. The algorithm for obtaining PPIs from orthologs is briefly described in Figure 3.5 below.



### Ortholog Prediction of Interaction Algorithm (OPIA)

We used a file from Integr8 containing all the orthologs of MTB strain CDC1551. Every file has two important segments (1) A file header containing the comment section which describes the contents of the file (2) The second section contains the ortholog information for 3773 proteins in the CDC1551 strain of MTB. The ortholog information for every protein is represented by a columnated line where columns represent information such as the species from which the orthologous protein is derived, the gene identity number, among

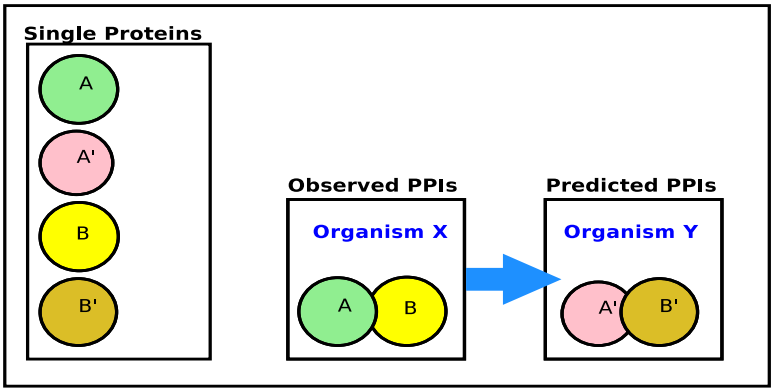


Figure 3.5: Illustration of the concept of interologs. The diagram illustrates the concept of interologs. For the proteins in the figure, A' and B' are homologs–more specifically orthologs of A and B respectively. Proteins A and B are observed to interact in organism X. Following this observation, the concept of interologs allows for the predictiction of the interaction between A' and B' in organism Y.

other information. In this section we are mostly concerned with two columns namely: (1) column containing the UniProt ID of the MTB protein, (2) a column containing the UniProt ID of an MTB protein's corresponding ortholog. OPIA algorithm is briefly outlined in the following steps.

1. Parse the ortholog file to retrieve all MTB-protein:Ortholog protein pairs.
2. Convert the list of the MTB-protein:ortholog pair into a dictionary structure such that each ortholog represents a dictionary key referencing an MTB protein for which it is an ortholog.
3. Parse the IntAct file and generate a list containing elements such that every element represents a pair of proteins reported in the the IntAct database to be interacting.
4. Go through each of the elements in the IntAct pairs list. For each element if both of its two component proteins are present as keys in the orthologs dictionary do the following:
   (a) Take the dictionary values (MTB proteins) corresponding to each protein constituting the IntAct pair.
   (b) Create an element constituting the MTB-proteins above and store the element in a list of inferred interacting proteins.

5. Repeat the process above for the entire IntAct database file.

Figure 3.6 shows a diagrammatic summary of the OPIA algorithm outlined above.

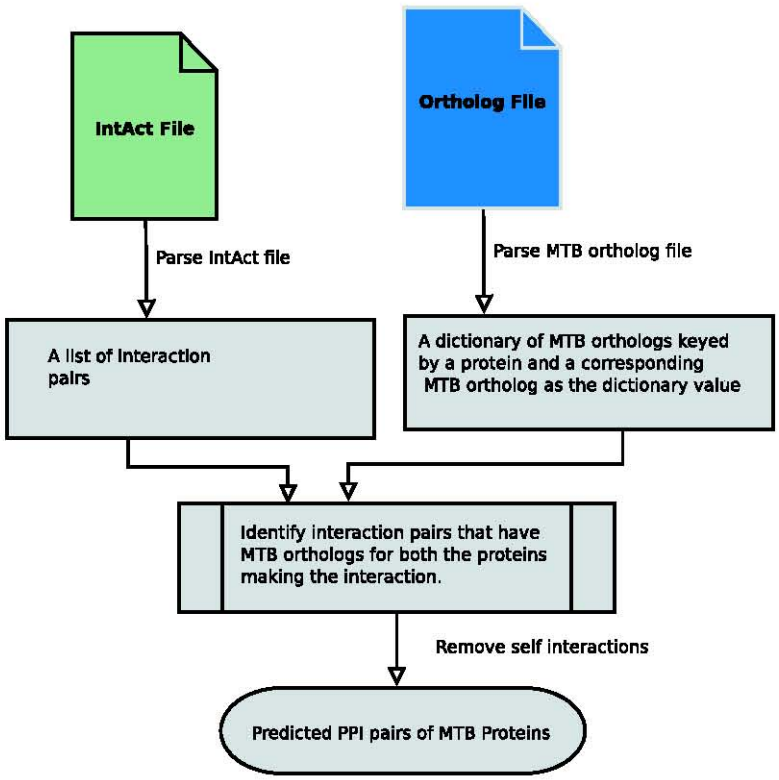


Figure 3.6: Illustration of Ortholog Prediction of Interaction Algorithm. A data flow diagram of the processing done on PPI file from the IntAct database and Ortholog file from http://www.ebi.ac.uk/integr8/EBI-Integr8-HomePage.do.

## 3.2 Integrating additional data

### 3.2.1 Functional Interactions–STRING

We obtained protein-protein interaction data from STRING (Search Tool for the Retrieval of Interacting Genes/Proteins), version 8.0, hosted at http://string.embl.de/ (June, 2009). The STRING database history can be obtained from the following website

http://string-db.org/server_versions.html. The current (as of November 2010) release of STRING is hosted at (http://string-db.org/). STRING is a database of known and predicted protein-protein interactions von Mering et al. (2007). Protein-protein interaction data in STRING are obtained from the following sources: 1) Genomic interaction data; 2) high-throughput experiments; 3) Co-expression data obtained from microarray experiments and 4) protein-protein interaction data mined from published articles in the PubMed database and those interactions deposited at the Munich Information Center for Protein Sequences (MIPS), as well some other small scale experiments mentioned in the release article cited above. A brief description of the data sources is provided below.

Genomic interaction data is based on the idea that functionally associated proteins are encoded by genes that experience similar selection pressures– the genes need to be maintained together, and regulated together such that the spatio-temporal interaction of the encoded proteins can be maintained in the cell von Mering et al. (2003b). The data binned in this category include data from gene fusions, genes that are always co-located, and those genes that share similar phylogenetic profiles across different genomes. The three methods above are explained in detail in the literature review in Chapter 2.

High-throughput experimental data is derived from experiments such as yeast-two hybrid (Y2H) and other experiments that characterize genes on a large scale. Co-expression data is derived from microarray experiments, and lastly PPI information is mined from articles published in PubMed to further enrich the collection of interaction data.

We queried the STRING database with all proteins in the genome of MTB (CDC1551 strain) and obtained all the interacting partners for each and every protein.

Every functional association in STRING has an attached interaction confidence score, which is evaluated on a scale ranging from 0–1.0.



### Computing STRING interaction confidence scores

Below is an excerpt from the blog that is maintained by the developers of the STRING resource on how the scores of string data are computed. The entire blog post can be read at http://string-stitch.blogspot.com/search?updated-min=2008-01-01T00%3A00%3A00%2B01%3A00&updated-max=2009-01-01T00%3A00%3A00%2B01%3A00&max-results=12.

A publication by the STRING developers von Mering et al. (2005), also briefly explains the probabilistic integrative technique employed in computing these scores.

The main concept that is discussed in the scoring scheme mentioned above, is that different methods used for identifying the STRING functional associations are weighted differently depending on the supporting evidence (publications) and Kyoto Encyclopaedia of Genes and Genomes (KEGG) database benchmark. A final score for PPIs that are identified by different independent methods is then computed in a probabilistic integration manner, described by the following equation:

$$totalscore = 1 - \prod_{i=1}^{n}(1 - score_i) \tag{3.3}$$

Where, $n$ is the number of independent evidences/methods supporting a particular PPI, and $score_i$ is the score of the $i^{th}$ evidence.



### 3.2.2 Subcellular localization prediction

As mentioned in Section 2.1.2, subcellular localization helps one identify where in the cell a protein is localized. Localization information of two proteins predicted to interact provides vital clues as to whether the two proteins are likely to interact *in vivo*. We use subcellular localization information to analyse the plausibility of our predicted interactions.

In this study, we employed two strategies to obtain protein subcellular localization data for MTB proteins, we obtained localization information from the UniProt database and predicted subcellular localization using the algorithm implemented in the PSORTb version 3.0 program. For some proteins subcellular localization information could not be obtained by either of these two approaches.

#### Subcellular localization from UniProt file

The comment section of a UniProt file contains information on a protein's subcellular location (for those proteins that have subcellular localization annotation). We parsed the UniProt MTB file using a locally developed script in Python programming language in

which we implemented regular expressions to extract the subcellular localization where available.

We stored the subcellular localization of the proteins in a dictionary with the proteins as dictionary keys and their corresponding subcellular localization values obtained from the UniProt file as dictionary values.

#### Subcellular localization calculated using PSORTb version 3.0

We downloaded a stand alone version of the PSORTb 3.0 program from (http://www.psort.org/downloads/index.html). The stand alone version is a library of programs implemented as PERL modules. The web implementation is available at (http://www.psort.org/psortb/).

After installing the stand alone version, we ran the program with the following options:

```
$psort -p cdc1551.fasta > output.txt
```

Where `cdc1551.fasta` is the input file containing the MTB proteins and `output.txt` is the file in which the output from running psort is redirected.

The output file contains, in the first column, an MTB protein and where possible the corresponding subcellular localization in the second column, otherwise the protein has `Unknown` written if the subcellular localization could not be predicted by the PSORTb version 3.0 algorithm.

Again, using a python script, we parse the result file retrieving subcellular localization for each protein and storing information in a dictionary where the MTB proteins are dictionary keys and their corresponding subcellular localizations are the dictionary values.

### 3.2.3 Gene Ontology (GO) data

An ontology Ashburner et al. (2000) is a formal representation of knowledge in a specific domain as a set of relationships that exist among objects that are members of that domain Barrell et al. (2009), Ashburner et al. (2000). The Gene Ontology project (GO) Ashburner et al. (2000), was initiated from the need to have a universal system for describing and

querying the function of genes. The GO vocabulary consists of three ontologies separated in the following categories: molecular function (MF); biological process (BP); and cellular component (CC). Each vocabulary that falls in any of the three ontologies above is structured as a directed acyclic graph (DAG, see Figure 3.7), wherein any term may have more than one parent as well as zero, one, or more children Huntley et al. (2009), Barrell et al. (2009). The GO terms form the nodes whereas the edges of the graph are represented as relations either of the form 'is-a' or 'part-of'. A *is-a* B means that A is a subclass of B. C *part-of* D means that whenever C is present, it is always part of D, but C need not always be present Wang et al. (2007).

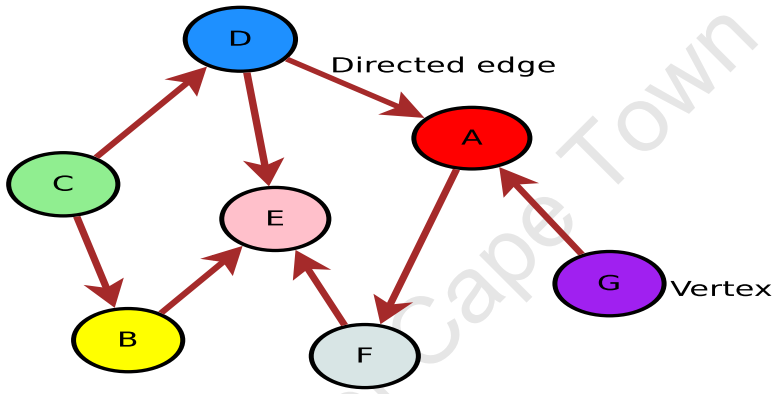


Figure 3.7: An example of a Directed Acyclic Graph (DAG). The graph is such that there are no directed cycles i.e there is no way that a path that starts at a vertex v eventually loops back to the same vertex. However, a vertex can have many paths leading into it, as well as many paths leading out of it. The relationships between the nodes of the graphs are depicted by the edges. The relations are either 'is-a' or 'part-of'.



Functional similarity has been found to be one of the best predictors or validators of protein-protein interactions Schlicker and Albrecht (2008). It then follows that, one way to predict or validate that two proteins interact would be to catalogue all the GO terms annotated to each of these proteins and find the functional similarity between the GO terms. Proteins found to share a substantial similarity in the three ontologies of biological process, molecular function and cellular component are likely to be involved in the same bio-chemical pathways, perform similar functions and be located in the same compartment in the cell, respectively. Substantial similarity in GO annotations found between predicted interacting proteins was used here to elevate and add confidence to the plausibility of our predicted interactions.

### 3.2.4 Slimming down GO terms

GO Slim is a summarized (slimmed) version of the GO vocabulary Biswas et al. (2002). In order to slim each ontology, a set of high-level terms has been created to cover most aspects of each ontology preferably without overlapping in paths of the GO hierarchy Biswas et al. (2002). These terms that give a more general description of a GO path are referred to as the GO Slim terms (http://www.geneontology.org/GO.slims.shtml). Basically, what the slimming process does is to collapse the more specific GO slim terms to their respective parent GO terms which are more representative. This process is summarized in the Figure 3.8 below.

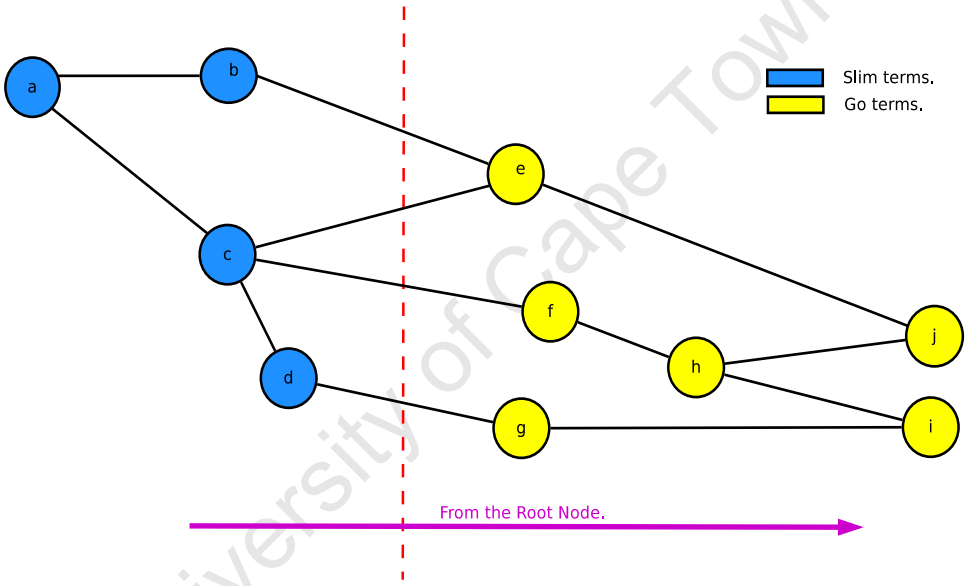




Figure 3.8: Illustration of the GO Slimming process. GO Slim terms are used to give a more general outlook of the GO graph by mapping terms in the child nodes onto the more representative parent nodes, hence the GO graph is 'slimmed'. From the figure above GO term e maps onto Slim terms b and c. GO terms f and h both map onto the Slim term c. GO terms i and j each has two paths leading directly into them from the from the Slim terms hence both have two slim terms (c,d) and (b,c) respectively.

In this study we use the GO Slim terms in each of the three ontologies: biological process; molecular function and cellular component in the following ways. We use the cellular component terms to enrich our sub cellular localization data by providing subcellular localization annotation to those proteins whose localization information could not be found by either running the PSORT program or using UniProt annotation. We use molecular

function ontology data for functional annotation of the predicted interacting protein pairs in a bid to identify the most highly represented functions among proteins that exhibit high prediction scores. Lastly, for biological process we use the slim terms to identify highly represented biological processes with the aim of identifying pathways that these interacting proteins are involved in. These processes are however hampered by the fact that not many of the MTB proteins have actually been annotated to GO terms, about 57% according http://www.ark.in-berlin.de/Site/MTB-GOA.html (October, 2011).

## 3.3 Protein-Protein Interactions from experiments

We also added to the list of PPIs predicted by the various methods, a small set (53 PPIs) of experimentally identified PPIs from MTB. We obtained this data from public PPI databases, such as the European Bioinformatics Institute (EBI) IntAct database. Note that this list included only PPI pairs that were composed of distinct proteins (self interactions were eliminated).



## 3.4 Scoring interaction confidence

We have developed two algorithms that we employed in determining PPIs. In addition, we have added data derived from functional linkages obtained from the STRING database as well as interaction data from experiments. Some of the predicted interactions have been observed to have been replicated across different prediction methods.

We propose a scoring scheme which we use to evaluate the confidence with which we attach to the predicted interactions. The scoring scheme takes into account the following issues:

- The type of method predicting a particular interaction, the methods are weighted differently.
- The number of methods supporting a particular interaction. In other words we use the number of independent methods supporting a particular interaction as evidence for that interaction. Hence our scoring scheme incorporates this observation by integrating the scores obtained from multiple methods.

### 3.4.1 Weighting the prediction methods

As mentioned above, different prediction methods are weighted differently with regard to the confidence that is attached to the likelihood that a prediction obtained from the method is true.

- **STRING** methods- We use the row final scores obtained from the STRING functional interaction pairs. The computation of the final scores is explained in von Mering et al. (2005). STRING considers the KEGG pathway support in calculating the score for a particular interaction. Simply, a PPI is considered to be of high confidence if the predicted interacting proteins belong to the same KEGG pathway.
- **Domain-Domain Interaction- with PDB evidence**- For domain-domain interactions that are supported by PDB evidence we set weight of 0.75.
- **Domain-Domain Interaction- with no PDB evidence**- In this category, we included PPIs also derived from DDIs, but in this case we did not have information that supported their interaction at structural level. For this category we set the score at 0.7.
- **Orthologs**- For PPIs obtained from ortholog transfer, we set the score at 0.75.
- **Experiment**- PPIs obtained from experiments were assigned a score of 1.0, which is the highest score possible since we deemed them to be of very high quality.



### 3.4.2 Computing the final score

For interactions that have been predicted by only one method, the weighting value of that method constitutes the final score, whereas, for PPIs that are predicted by more than one method, the weight evidence of each method contributes to the final score. The different methods/evidences are assumed to be independent of each other. Therefore, the weight contribution of each method is added in an integrative manner as explained in Equation(3.3).

Where $score_i$ denotes the score for method $i$, and $i$ ranges from 1 to $n$. In this case, $n$=4, which is the total number of independent methods considered in this study. The weighting

is adapted from the scoring scheme described in von Mering et al. (2005) for scoring PPIs in STRING von Mering et al. (2007).

From Equation(3.3) , the contribution of different scores are treated independently hence, for the addition in an integrative fashion, the different score contributions are multiplied. Note that the scores are not averaged since the averaging process will bring down the effects of methods that are weighted highly when taken together with low confidence methods. Take an hypothetical example of PPI supported by two methods with confidence scores of 0.1 and 0.9 respectively. By simply averaging the scores, a total score of 0.5 is obtained, as opposed to a total score of 0.91, that is obtained when the scoring function in Equation(3.3) is applied.

## 3.5 Brief analysis of the generated network

It has been mentioned in text that proteins perform their functions through the intricate interactions that they form with each other (see 2.2). In this analysis, we use node centrality, a widely used concept in both graph theory and network analysis Freeman (1978). Basically, the centrality of a node simply expresses how crucial the node is in relation to the rest of the network. Various studies have suggested that a node centrality directly correlates with its biological importance Pang et al. (2010), Zotenko et al. (2008), He and Zhang (2006). In other wards, crucial nodes tend to have more connections than less crucial ones.

Three quantities namely: closeness, degree and betweenness are used to calculate a node's centrality. In this study we performed an analysis, that is similar in set up to the one in http://www.babelgraph.org/wp/?p=1 that sought to find out which members of a given social network were more connected to the rest of the group than other members of that social network. The algorithms were implemented in the form of an R-Language package (igraph). The result of the analysis is a plot showing the relative centrality of the nodes in Figure 4.7

## 3.6 Evaluating evolutionary relationship

### 3.6.1 Calculating dN/dS

The basis of dN/dS, otherwise known as a phylogenetic analysis based method for detecting molecular evolution, starts with first defining a multiple sequence alignment (MSA) of orthologous genes from the organisms to be aligned Chenna et al. (2003). We consider 3 organisms within the MTB family for our analysis. The three organisms are: MTB (KZN strain); *Mycobacterium bovis* (strain AF2122/97) and *Mycobacterium tuberculosis* (strain Oshkosh/CDC1551), the MTB strain that most of the analysis in this study was based on. We only use 3 organisms for the dN/dS calculation since the number of CDC1551 proteins with orthologs in all the of the organisms dropped.
*Procedure*

1. First we downloaded the gene set files of the 3 organisms above from (http://www.ebi.ac.uk/integr8/EBI-Integr8-HomePage.do)

2. We also downloaded the MTB (strain Oshkosh/CDC1551) ortholog file from the same site.

3. We then identified all the MTB (strain Oshkosh/CDC1551) orthologs in the two other organisms, ending up with a tab separated list that has in its first column an accession number of MTB (strain Oshkosh) protein, and the next two columns having accession numbers of its corresponding orthologs in MTB (KZN strain) and *Mycobacterium bovis*, respectively.

4. For every row in the file in (3) above, we retrieved sequences in FASTA format for all the accession numbers in the order that they appear in the file starting from the MTB CDC1551 strain. We saved the three sequences in a file aptly assigned the name of the MTB CDC1551 protein. At this stage the sequences are ready for MSA.

5. We perform MSA and dN/dS calculation using the HyPhy package Pond et al. (2005). The process can be broken down into the following sub-processes implemented as wrapper files in the HyPhy batch language (HBL). There are different wrapper files that perform the following actions

(a) MSA using the in-built alignment program in HyPhy.

(b) Estimate the phylogenetic tree of the three sequences using the Neighbour Joining (NJ) algorithm Saitou and Nei (1987) for phylogenetic tree reconstruction, again implemented in the HyPhy package.

(c) Calculate dN/dS using the tree and the alignment file from the processes (a and b) above.

(d) The dN/dS values for each alignment file is stored under the alignment file name–this represents the dN/dS value for the protein represented by the file name.

### 3.6.2 Codon volatility

Codon volatility as a method to estimate relative selection pressures was developed by Plotkin in Plotkin et al. (2004). The core component of this method is the observation that, if a protein coding region of a nucleotide sequence has undergone an excess number of amino-acid substitutions, then the region will, on average, contain an overabundance of 'volatile' codons, in comparison to the genome as a whole. In this study, we employ the definition used in Plotkin et al. (2004) which states that codon volatility is the probability that the most recent mutation at a codon site that gave rise to the observed codon resulted in an amino acid change. Basically, the probabilistic measure identifies the number of ancestral codons (stop codons not taken into account), that would, with a single point mutation, give rise to the observed codon. The higher the probability of giving rise to an amino acid change, the higher the volatility and *vice versa*.

The main advantage of the codon volatility method over the widely used phylogenetic analysis method is that, unlike the latter, it is not affected by poor MSAs, according to Codoner and Fares (2008). Another advantage is that it is possible to perform an analysis using only a single genome sequence hence reducing problems such as lack of clear orthologs that hamper phylogenetic/tree-based methods.

The codon volatility algorithm is implemented as an online accessible program at `http://mathbio.sas.upenn.edu/volatility/cgi-bin/volatility.pl`. The details of how the algorithm is implemented are in the main body of Plotkin et al. (2004), and its evaluation with comparative genomics methods is available in the supplementary section of this paper.

We applied the algorithm to the MTB CDC1551 proteins.

### 3.6.3 Background to functional similarity calculation

Several methods have been proposed for calculating functional similarity between genes/proteins. Most of these methods borrow heavily from approaches used in information theory (IT).

One of the simplest, but least sensitive of these methods involves counting the number of GO terms that overlap between the two proteins, otherwise known as the term overlap (TO) Mistry and Pavlidis (2008). The Term Overlap (TO) score is evaluated as follows:

$$TO(P_1, P_2) = \sum (GA_1 \cap GB_2) \tag{3.4}$$

Where $P_1$ and $P_2$ are two proteins and $GA_1$ and $GB_2$ are GO terms directly annotated to the two proteins (including their parent GO terms). Another approach would be to apply Jaccard's coefficient Ivchenko and Honov (1998) to calculate the similarity score between two proteins predicted to interact. This approach finds the ratio of shared GO terms to that of the union of all the GO terms annotated to the two proteins, while eliminating redundancies. This expression of the similarity score between two proteins A and B is as follows.

$$Sim_{A,B} = \frac{G_A \cap G_B}{G_A \cup G_B} \tag{3.5}$$

Where $G_A$ and $G_B$ are the vectors of GO terms annotated to Protein A and B respectively. The similarity between the two proteins (A and B) above, is considered to increase as the quotient approaches the value 1. The two proteins are considered identical if the quotient equals 1 and it indicates that the two proteins share all their GO terms.

### 3.6.4 Calculating similarity between GO terms

Functional similarity scores for protein pairs have been used in various spheres of functional genomic research including finding functional clusters of proteins, predicting protein functions, and for predicting protein-protein interactions Wang et al. (2010).

GO terms associated with a protein provide an annotation as to what biological process

the protein is involved in, what molecular function the protein performs and lastly, the cellular compartment in which a protein is located. Pairwise GO similarity scores between proteins predicted to interact can be calculated over the three ontologies (biological process, molecular function and cellular component). The scores thus generated can be used to validate predicted interactions. For instance, in theory one would expect interacting proteins to be located in close proximity to each other Bhardwaj and Lu (2005), Dandekar et al. (1998). Therefore, the GO similarity score for the cellular component ontology calculated between between two interacting proteins is expected to be high. Likewise, for the molecular function and biological process ontologies, interacting proteins more often than not would be expected to perform similar functions Fraser et al. (2004), and be involved in similar biological processes. Again, one would expect high scores for pairwise similarity scores for predicted protein interactions in these three ontologies (MF, BP, CC).

We calculate the semantic/functional similarity scores between the GO terms annotated to any pair of predicted interacting proteins. We propose that, any pair of predicted interacting proteins which show high similarity in their GO terms in the three ontologies are more likely to be true interactions.



**Information Content based similarity measurement**

Resnik first proposed an information content (IC) based method for inferring semantic similarity Resnik (1995). The main component of the IC based similarity measurement, is the frequency of occurrence of a term. Terms that occur ubiquitously carry less weight (information) than terms that rarely occur in an annotated data set. Most approaches to scoring semantic similarities between GO terms developed thus far rely on the annotations provided in the GO databases Sevilla et al. (2005), Jain and Bader (2010). The information content (IC) that is based on annotation is obtained by the following equations:

$$freq(t_1) = annot(t_1) + \sum_{c \in children(t_1)} freq(c) \tag{3.6}$$

Where $freq(n)$ is the frequency of the term n. The probability of a term $t_1$ is thus defined as:

$$prob = \frac{freq(t_1)}{freq(root)} \tag{3.7}$$

Information content of a term, $IC(t_1)$ is then given by:

$$IC(t_1) = -log(prob(t_1)) \tag{3.8}$$

From the equation (3.9) above, terms that occur ubiquitously equate to less information content than terms that occur rarely.

### Topological position based IC

In this work we use a method developed in our group Mazandu and Mulder (2011). At the time of writing this thesis (February, 2011), the paper describing the method is still under review. The method uses the topological characteristic of a GO term to infer its information content instead of the inaccurate frequency of annotation based methods. The GO database is constantly updated with new information therefore similarities calculated based on frequency of annotation do not always provide consistent information.

*Definition 1*
$\mathcal{T}_{\mathcal{GO}}$ is the set of GO terms. $[x, y] \in \mathcal{T}_{GO}$ depicts the relationship that $x$ is lower than $y$ in the GO DAG.

$\mathcal{L}_{GO}$ is a set of links in the GO DAG such that, $(a, b) \in \mathcal{L}_{GO}$ is a link association between parent term $a$ and child term $b$.

The algorithm defines the topological position characteristic of a term $t$ as

$$\mu(t) = \begin{cases} 1 & \text{if } t \text{ is root} \\ \prod_{x \in \mathcal{P}_t} \frac{\mu(x)}{C_x} & otherwise \end{cases} \tag{3.9}$$

Where $\mu(t)$ is a topological position characteristic of $t$. This value is recursively obtained using the parents of $t$ obtained from the set $\mathcal{P}_t = x : (x, t) \in \mathcal{L}_{GO}$. $C_x$ is the number of

children of the term $t$.

A topological position is a function $\mu : \mathcal{T}_{\mathcal{GO}} \longrightarrow [0, 1]$, such that for any term $i \in \mathcal{T}_{\mathcal{GO}}, \mu(t)$ defines the reachability measure of an instance of term t. From the above definitions, it follows that $\mu$ is a monotonically increasing function as one moves towards the root node, with the root node having the maximum reachability which equals 1.

*Definition 2*

Let $[x, y] \in \mathcal{T}_{\mathcal{GO}}$ x and y are topologically synonymous, denoted as $x = y$, if the following properties are satisfied.

- $IC_T(x) = IC_T(y) or \mu(x) = \mu(y)$
- There exists one path $\mathcal{P}_{xy}$from x to y.

Following the satisfaction of the two properties above, two GO terms are considered equal if and only if they are either the same or topologically identical terms.

Suppose there exists a path $p_{xy}$ from the term x to the term y, it follows that y is a more specific term when compared to x or x is more general in comparison to y. This can be expressed as follows:

$$x <^{GO} y if IC_T(x) < IC_T(y) or \mu(y) < \mu(x) \tag{3.10}$$

In order to define the closeness between two GO terms, we define the topological position $\mu_s(x, y)$ of $x$ and $y$ as that of their common ancestor with the smallest topological position characteristic, i.e.,

$$\mu_s(x, y) = min\mu(t) : t \in \mathcal{A}(x, y) \tag{3.11}$$

Where $\mathcal{A}(x, y)$ is the set of ancestral terms shared between $x$ and $y$.

We now define the semantic similarity between two GO terms thus:

$$S_{GO}(x,y) = \frac{IC_T(x,y)}{max\{IC_T(x), IC_T(y)\}} \tag{3.12}$$

Where $IC_T(x,y) = -ln\mu(x,y)$, the topological information shared by the two terms $x$ and $y$. Here is an example where you need large space in expression:

$$S_{\mathcal{F}}(p_1,p_2) = \frac{1}{2}\Big[\frac{1}{T^X_{GO}(p_1)}\sum_{t\in T^X_{GO}} S_{GO}(t,T^X_{GO}(p_2)) + \frac{1}{T^X_{GO}(p_2)}\sum_{t\in T^X_{GO}} S_{GO}(t,T^X_{GO}(p_1))\Big] \tag{3.13}$$

Where $S_{GO}(t,T^X_{GO}(p)) = 1 - d_{GO}(t,T^X_{GO}(p))$, with $d_{GO}(t,T^X_{GO}(p))$ as the distance between a term $t$ and the set of terms $T^X_{GO}(p)$ for a given protein $p$, defined as

$$d_{GO}(t,T^X_{GO}(p)) = min\{d_{GO}(t,s) : s \in T^X_{GO}(p)\} \tag{3.14}$$

Given that $d_{GO}(s,t) = 1 - \mathcal{S}_{GO}(t,s)$, we express:

$$\mathcal{S}_{GO}(t,T^X_{GO}(p)) = max\{\mathcal{S}_{GO}(t,s) : s \in T^X_{GO}(p)\} \tag{3.15}$$

We used Equation (3.15) derived above to calculate the similarity in GO annotations closeness between pairs of our predicted PPIs in the three ontologies. We then plotted the distribution of the similarities for the set of our predicted PPIs.

### 3.6.5 GO enrichment analysis

In GO enrichment analysis we are interested in investigating GO term representation. Specifically, we seek to determine GO terms that are either over-represented or under-represented in our set of predicted PPIs with respect to the rest of the genome. The knowledge that we derive from this analysis aids us in annotating the proteins in our interaction set. Our strategy for calculating GO enrichment involves running Blast2GO Conesa et al. (2005), a functional enrichment analysis tool which can be downloaded at http://www.blast2go.org/start_blast2go. Blast2GO uses Fisher's exact test Fisher

(1922), to calculate the GO enrichment for a given set of proteins. We ran the Blast2GO program on the PPIs that we consider to be of high confidence (all proteins total scores $\geq 0.9$).

The Blast2GO program also performs a direct GO term count of terms in an analysis. We included the results for the various ontologies in Figures 4.20, 4.21, 4.22, for the biological process, molecular function and cellular component ontologies respectively.

## 3.7 Deriving biological meaning

In this section we attempt to analyse biological significance from PPIs thus far predicted. We employ the gene ontology (GO) as the basis of our biological significance analysis. We hypothesize that interacting proteins are expected to share significant similarity in the three ontologies; biological process (BP), molecular function (MF) and cellular component (CC). We perform this analysis by: 1) first calculating functional similarity values for each pair of the predicted PPIs in the entire PPI network, and 2) performing gene set enrichment analysis with the aim of finding over represented functional classes in a bid to annotate the proteins in the interaction network.

# Chapter 4

# Results



In this chapter we present the results that we obtain from applying two algorithms to computationally identify interacting proteins within the genome of MTB. We also mined the STRING database von Mering et al. (2003a), for direct interactions and functional associations among MTB proteins and generated results for molecular evolution and biological evaluation of our predicted PPIs. We present these findings below.

## 4.1 Interaction Prediction Results

In this section we examine the results obtained from the algorithms that we proposed for inferring PPIs. We developed two algorithms (DEA, in Section 3.1.1, and OPIA, in Section 3.1.2), for inferring PPIs. The algorithms use the concept of known DDI interactions and orthologous interaction transfer (interologs) from other species to infer PPIs respectively. In addition, we retrieve, from the IntAct database PPIs that have been experimentally determined and add them to our protein interaction dataset. The experimentally determined interactions, at 53 distinct interaction pairs, are a relatively small list when compared with the interactions obtained by the other prediction methods. The highest number of PPIs predicted by a single method is 20261 interaction pairs, from functional interactions derived from the STRING database.

The following table shows a summary of the distribution of PPIs we obtained through the

different methods described in the previous chapter.

Table 4.1: Distribution of MTB PPIs by method.

| Method | No. of MTB PPIs | No. of unique proteins |
|---|---|---|
| Experiment | 53 | 35 |
| Domain Interactions PDB | 864 | 230 |
| IntAct Domain Interactions | 5115 | 199 |
| Functional Interactions (STRING) | 20261 | 3425 |
| IntAct Ortholog Interactions | 1702 | 331 |

The information in table 4.1 is graphically displayed in a bar chart in Figure 4.1.

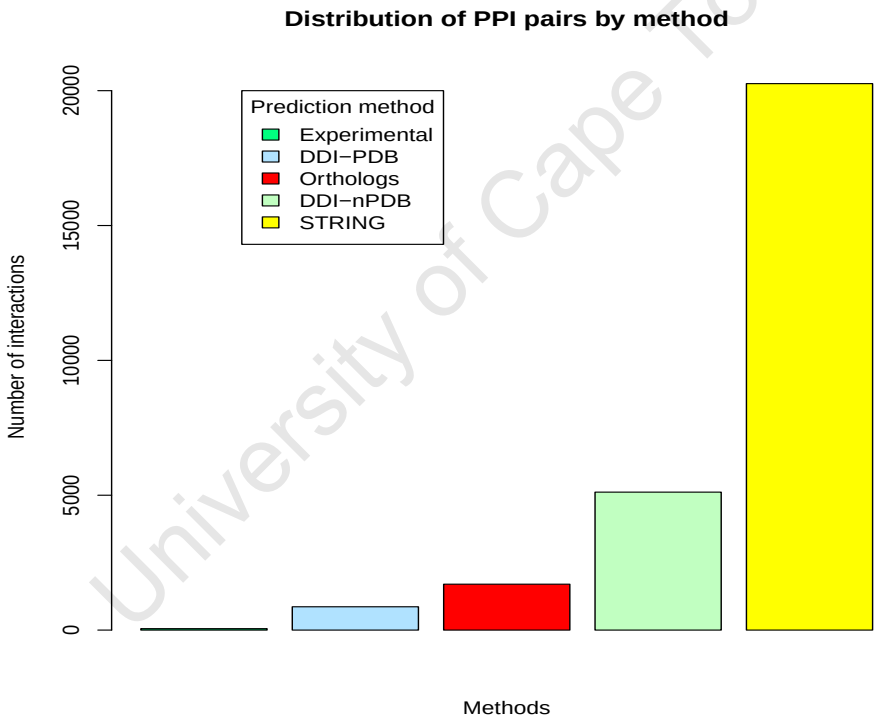




Figure 4.1: Barplot of distribution of interactions by method, showing the distribution of predicted interactions by method. The y-axis indicates the number of PPIs and the different interaction detection methods are color coded as shown in the graph legend.

As described in Section 3.2.1, STRING is a database of known and predicted protein-protein interactions von Mering et al. (2007; 2003a).

STRING maintains protein interaction data both from direct (physical) interactions and indirect (functional) associations. STRING integrates up to 8 different (treated independently) methods to infer PPIs. Given the wide coverage of STRING we propose that PPIs identified by our prediction algorithms in Chapter 3 and replicated in STRING suggest that they are of high confidence. We present the level of replication of interactions predicted by our algorithms and replicated in STRING below.

Table 4.2: Number of Protein-Protein Interactions replicated in STRING.

| **Method** | **No. of replications** | **Avg. STRING Score** |
|---|---|---|
| Domain Evidence Algorithm (DEA) | 88 | 0.61 |
| Domain Evidence Analysis PDB | 100 | 0.77 |
| Ortholog Prediction of Interaction Algorithm (OPIA) | 180 | 0.86 |



Table 4.2 shows the number of PPIs that we identified by OPIA and DEA that were also replicated in the STRING database. We interestingly found that the average STRING score for the replicated data correlated with the weighting values that we defined in 3.4.1. For example, domain-domain interactions with 3-dimensional structural information from the PDB, would in theory be expected to have higher scores on average than for the IntAct DDIs, for which we don't know whether they have structural support or not from Table 4.2 above.

We then investigated the level of replication that a particular predicted PPI has across the different prediction methods in Table 4.3.

Table 4.3: Distribution of interaction overlap across the methods. *Exp* is for PPIs with experimental evidence, *Dom1* is for IntAct DDIs for which we do not have structural support, *Dom2* is for DDIs with PDB structural support, STRING, is for PPIs derived from the STRING database, and *All* is for the aggregate of all these methods.

| | **Exp** | **Dom1** | **Dom2** | **IntAct-Orthologs** | **STRING** | **All** |
|---|---|---|---|---|---|---|
| Exp | 100.00 | 0.435 | 0.000 | 0.000 | 0.064 | 0.192 |
| Dom1 | | 100.00 | 0.518 | 0.078 | 0.473 | 3.314 |
| Dom2 | | | 100.00 | 0.220 | 0.347 | 18.55 |
| IntAct-Orthologs | | | | 100.00 | 0.820 | 6.174 |
| STRING | | | | | 100.00 | 73.49 |
| All | | | | | | 100.000 |

From the table, we show that the level of replication is not very high while taken proportionally across the databases, probably due to the challenges discussed in Section 2.2.4.

### 4.1.1 Number of interaction partners for proteins

Figures 4.2 and 4.3 show the distribution of the number of interaction partners that proteins have. Proteins in the high scoring region tend to have fewer interaction partners compared to the distribution taken for the whole PPI set.

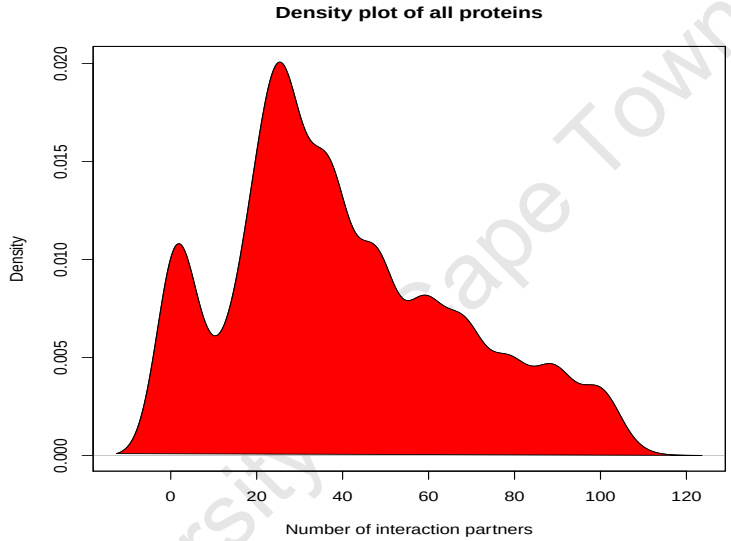




Figure 4.2: Density plot of the number of interaction partners, taken for the whole set of our predicted PPIs.

## 4.2 Confidence score results

In this section, we present the distribution of the total confidence scores for the PPIs, and give an explanation for the possible cause of the observed trend. Figure 4.4, shows the histogram that displays the distribution of the total scores across the network of the predicted PPIs.

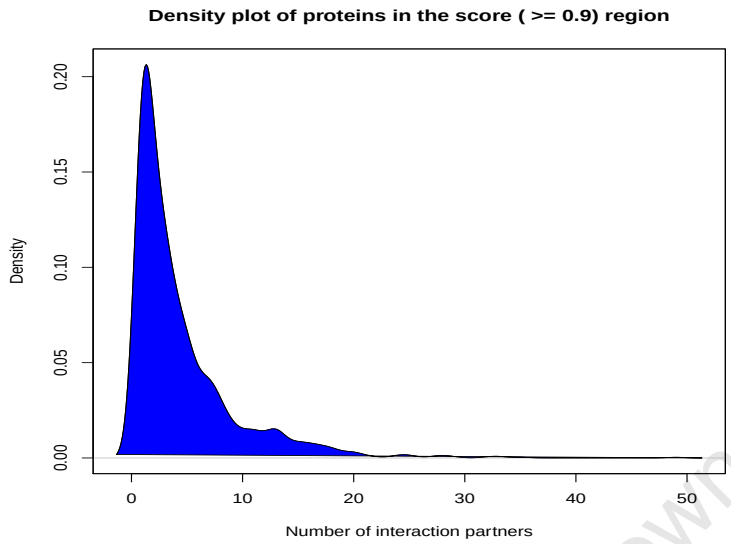


Figure 4.3: Density plot for proteins in the total confidence score range of 0.9–1.0. PPIs in this score range are considered to be of high confidence.

Fig 4.5, shows the results displayed in Figure 4.4 as a cumulative histogram. Note the sharp rises in bar lengths at score level 0.7 and close to 0.8 showing sudden increase in the number of PPIs as opposed to the steady incline witnessed in scores lower than 0.7 and higher than 0.8. This shows that the majority of PPIs fall in between these two scores.

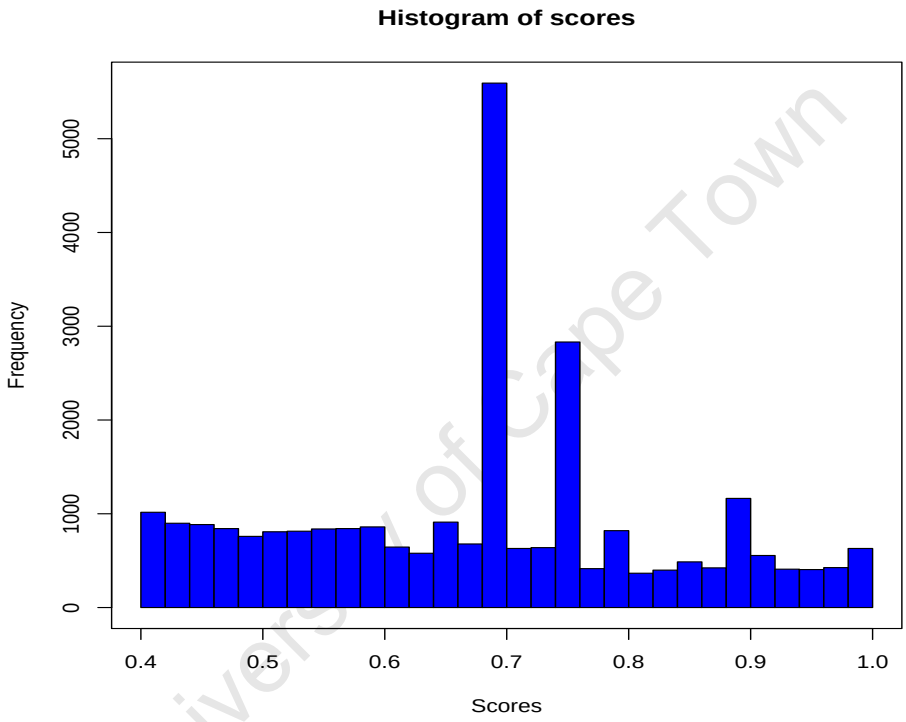




Figure 4.4: Distribution of protein-protein interaction confidence scores. The lowest score is maintained at 0.4 whereas the highest score observed is 1.0. From the graph above, it is clear that the bulk of the interactions score within the region of 0.7

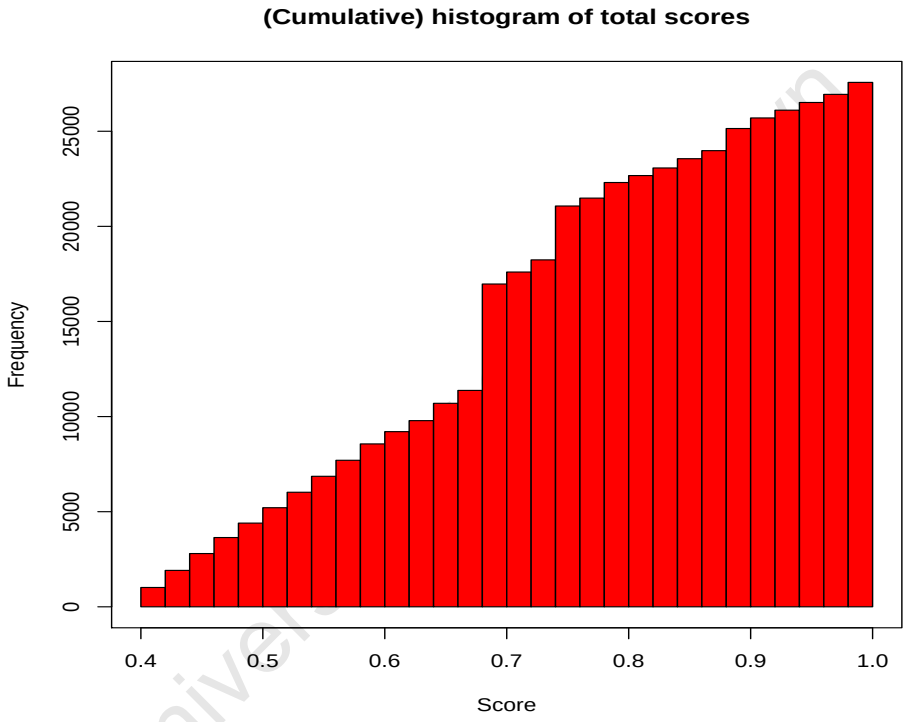


Figure 4.5: Cumulative histogram showing the distribution of PPI confidence scores. The majority of the scores fall in the range between 0.7 and 0.8

We postulated that confidence scores correlate with subcellular localization. We investigated whether PPIs that had high total scores also shared the same subcellular localization. Figure 4.6 shows the percentage of PPIs that share the same subcellular localization in every score category. Note the steady increase of the percentage of PPIs sharing the same subcellular localization as the score increases. This suggests a consistency in our PPI network with regards to the confidence score.

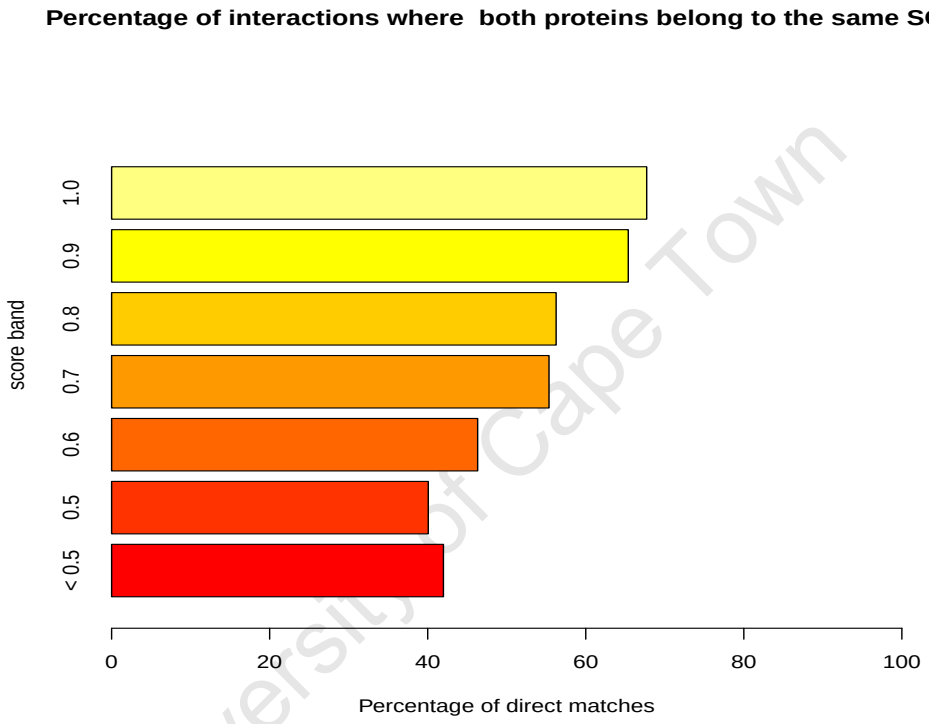




Figure 4.6: A barplot of percentage of protein pairs both in the same subcellular localization (SCL). The graph above shows the distribution of percentage of PPIs wherein both of the constituting proteins are predicted to be located in the same cellular compartment. The scores are computed for the different score bands. Each score band for total scores greater than 0.5 has a width of 0.1. The band 0.5 constitutes all of the score values less than or equal to 0.5.

Computational ways of predicting PPIs are met with challenges, some of which are mentioned in Section 2.2.4. These challenges when not handled correctly may lead to spurious predictions. We select interactions that we consider to be of high confidence to generate an example network in Figure 4.7.

We chose a total score of 0.7 to represent our cut off for medium to high confidence

interactions (refer to Section 3.4.2 for how the final score is computed). Note that the total score of 0.7 accommodates all the interactions predicted by our algorithms while at the same time not compromising too much on the raw scores obtained from STRING. In other words, a score 0.7 is still high even for interactions that are only predicted by STRING. There are a total of 15615 PPIs that have a score of 0.7 or more, considered to be the medium to high confidence subset.

## 4.3 Network analysis of high confidence interactions

We selected high confidence PPIs, those that had scores greater than 0.9 for network analysis.The results for network analysis that is displayed in Figure 4.7, show that there are a few high-degree nodes in the network evident from the comparative size of the nodes in addition to the varying intensity of darkness. The darker and bigger a node is, the more central it is to the network. Studies such as those by Zotenko et al. (2008), Pang et al. (2010), He and Zhang (2006) have suggested that topological prominence of a protein in a network can be a good indicator of biological importance. We did not do a detailed network analysis as this wasn't one of the main objectives of the study but the plot in Figure 4.7 suggests that this is one of the possible areas in which this study can be extended.

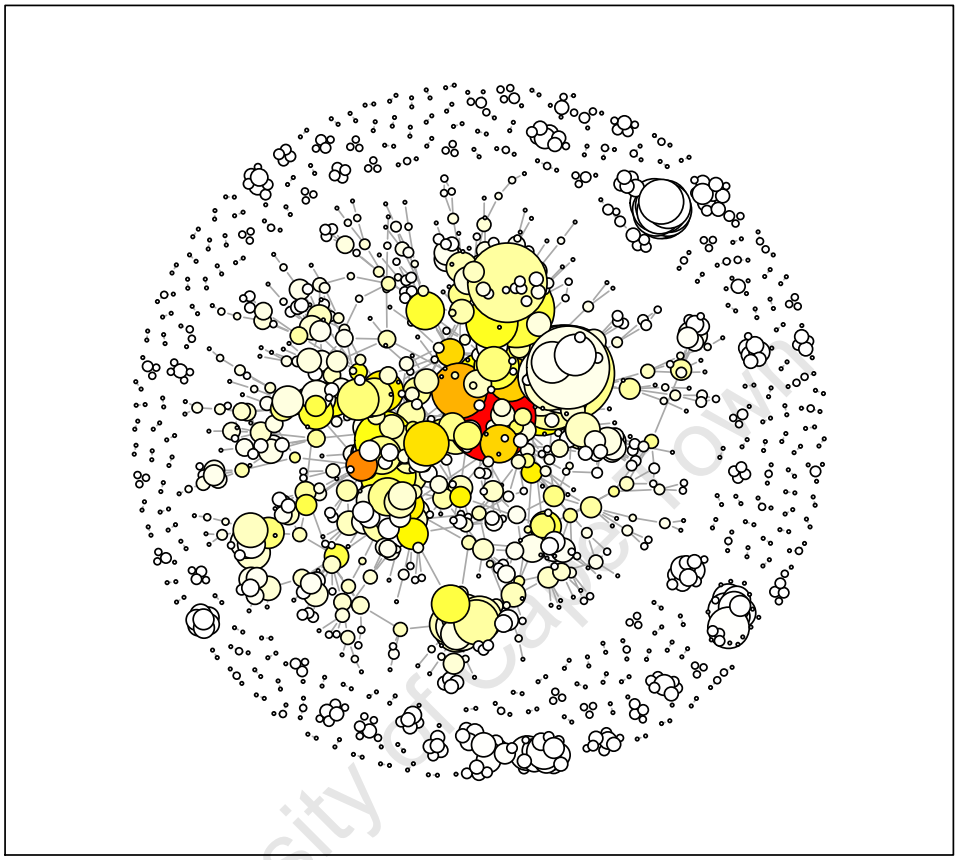



Figure 4.7: Protein-protein interaction network betweenness centrality for high confidence interactions. The size of the nodes are scaled according to the degree of the node i.e. highly connected nodes are bigger in size than ones that are less connected. Betweenness centrality of a vertex in graph analysis is the number of shortest paths on the network that pass through the vertex. A high betweenness score indicates that the vertex acts as a mediator of connections betweeen other vertices. Here the betweenness scores were mapped to a heat color scheme in with the color red indicating high betweenness scores.

## 4.4 Evolutionary relationship results

In Section 3.6, we discussed two approaches to identifying molecular evolutionary dynamics. In sections 3.6.1 and 3.6.2 we discussed the calculation of dN/dS and codon volatility respectively. Here we present results obtained from these two approaches.

### 4.4.1 Distribution of dN/dS and comparison to codon volatility

We discussed calculation of dN/dS in Section 3.6.1. From the results, most of the dN/dS values are clustered around 0. This clustering around 0 indicates that most of the proteins, which have a value other than 0, are under purifying selection (see the interpretation of dN/dS calculation in Section 2.3.1). Indeed most of the protein sequences were conserved across the MTB orthologs that we used in this study. There is also a substantial clustering, although to a limited scale of dN/dS values, significantly greater than 1 (indicative of positive selection acting on the proteins in this category) evident from the lower peak in Figure 4.8 which shows the density plot of dN/dS values.



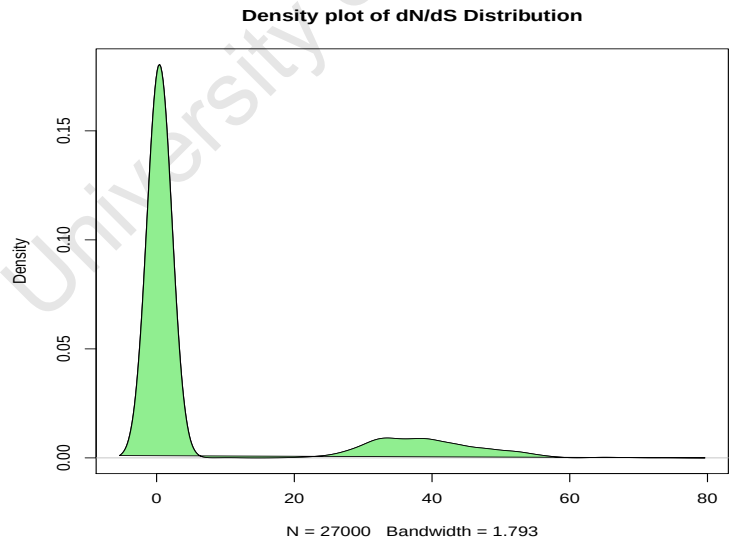


Figure 4.8: Density plot of dN/dS. A density plot showing the distribution of dN/dS values. Note the two peaks, the first one for values around 0, and the second one for values around 40, indicating purifying selection and positive selection respectively for the peaks.

We isolated PPIs for which both of the proteins in the interaction pairs had dN/dS values greater than 1. We found a total of 630 PPIs in this category. There are also 19868 PPIs where both partners constituting the PPI have dN/dS value less than 1 (purifying selection). Following the definition of codon volatility in Section 3.6.2, in theory one would expect low codon volatility values to be associated with high dN/dS values (positive selection). This in effect will impress a negative correlation coefficient upon the correlation between codon volatility values and dN/dS values. Indeed, a negative correlation (-0.1489960, P-Value 2.2e-16) is obtained between these values. Figure 4.9 shows the plot of codon volatility vs dN/dS. Note that due to the conservation of protein sequences across the several species under investigation, most dN/dS calculations were undefined ($dN = 0$ and $dS = 0$), thus set to the same random number, which happens to be less than 1 hence counting as negative selection going by the interpretation of dN/dS calculations.

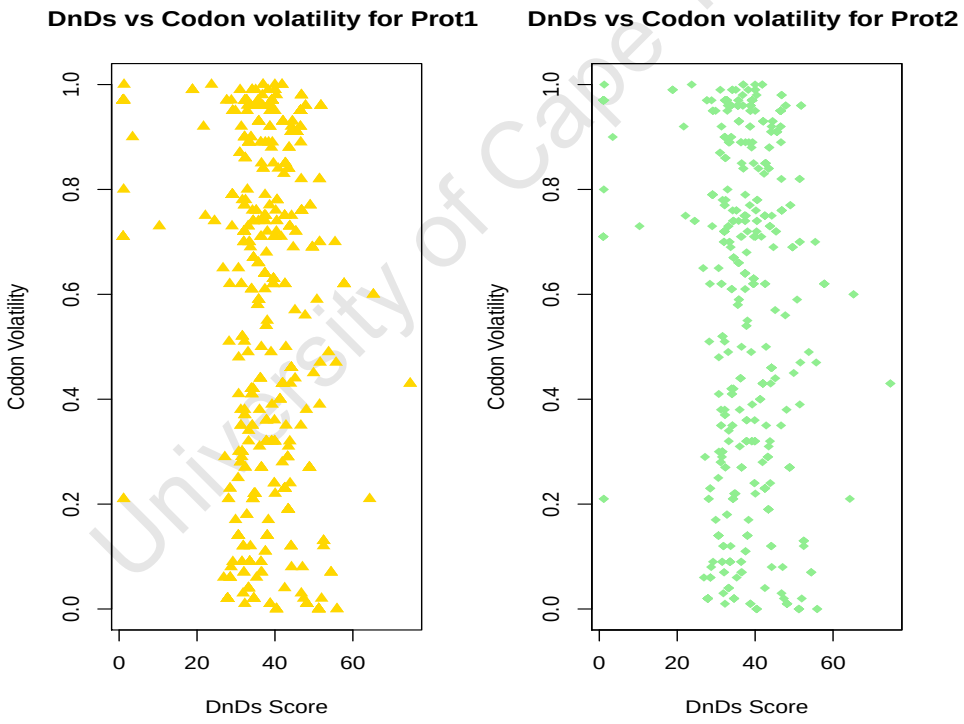


Figure 4.9: A plot of dN/dS vs codon volatility. A Pearson correlation value of -0.1489960 (P-Value 2.2e-16) is obtained when a correlation test is performed on codon volatility score for the *Protein 1* against dN/dS value for *Protein 1*. Likewise, a correlation test on *Protein 2* codon volatility score and dN/dS score yields -0.1283291. Where *Protein 1* and *Protein 2* constitute an interaction pair.

### 4.4.2 Correlation between interacting proteins

We have calculated Pearson's correlation coefficient for codon volatility values of the pairs of predicted PPIs. Overall, there is a weak positive correlation of (0.0660718) at a significance level of 2.2e-16 between the predicted interacting pairs. We then tried to determine whether predicted PPIs of different interaction scores correlate differently. We present Pearson's correlation coefficient values calculated for total interaction scores in different bands. In other words, we band the score range (0-1) into different bands, and for each band we calculate the Pearson's Correlation Coefficient value between dN/dS values and corresponding codon volatility values.

Table 4.4 below, shows the Pearson's correlation value in the different bands. This relationship is represented graphically in Figure 4.10.

Table 4.4: Distribution of interaction overlap across scores.

| **Score range** | **Correlation Value** |
|---|---|
| Greater than 0.9 | 0.1041893 |
| Greater than 0.8 | 0.09280075 |
| Greater than 0.7 | 0.07081366 |
| Greater than 0.6 | 0.06518539 |
| Greater than 0.5 | 0.06630598 |
| Greater than 0.4 | 0.0660718 |
| Greater than 0.2 | 0.0660718 |
| Greater than 0.1 | 0.0660718 |



Figure 4.11 shows a logarithm plot of the number of PPIs that are predicted to undergo the same selection pressure varied over different score ranges. From the plot, the trend is similar for both proteins undergoing negative selection and those that undergo positive selection. However, it is also clear from the graph that for most PPIs, the proteins are under purifying selection (the line plot corresponding to negative selection is on top of the one that corresponds to positive selection). Again PPIs in 0.7-0.8 confidence score range lead in frequency in either positive or negative selection.

Figure 4.12 shows a plot of dN/dS values drawn from pairwise interactions of the entire interaction set. There are protein interactions between proteins with low dN/dS values

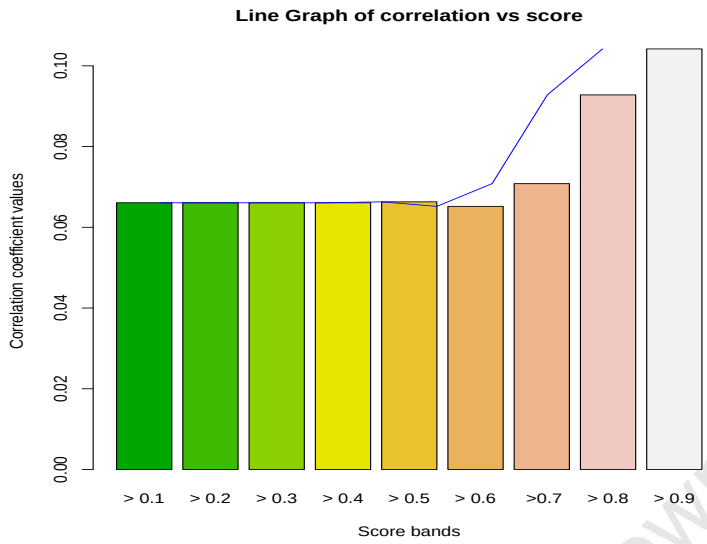


Figure 4.10: A plot of Pearson's correlation coefficients by score, as calculated for different score ranges.

interacting with proteins with significantly elevated dN/dS values. This is represented in the graph by clustering along both the $x$, and the $y$ axes. There is also a cluster of interactions involving proteins with high ($> 1.0$) dN/dS values. This is shown by the clustering of data points at the center of the graph.

Shown in Figure 4.13 is a plot of randomly generated pairs of proteins. The graph is almost an identical image of the plot in Figure 4.12, suggesting that there is no real evidence for a correlation in dN/dS values for interacting proteins. However, this may because 4.12 includes interactions of all scores, although similar observation is found when we look at the plot for high confidence interactions in Figure 4.14

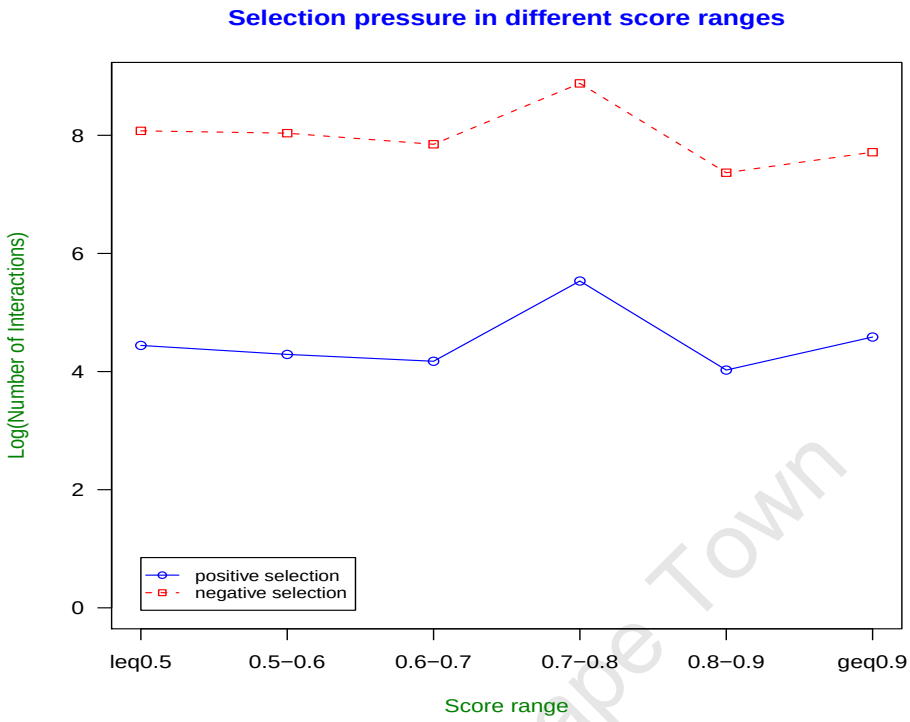


Figure 4.11: A Log plot of number of PPIs under either negative or positive selection. The blue line plot denotes the log(N) plot of the number of proteins undergoing negative selection whereas the red line plot denotes the log(N) plot of the number of proteins undergoing positive selection.



## 4.5 Biological interpretation results

Our strategy for biological relevance evaluation relies on the measures that we proposed in Section 3.7. We suggested that gene set enrichment analysis plus similarity calculation across the ontologies could aid in functional annotation of our predicted PPIs. We relied upon the hypothesis that interacting proteins would be expected to share substantial similarity across the three ontologies (MF, BP, CC). We also set out to find the functional categories and biological processes that were over-represented in our predicted PPI set.

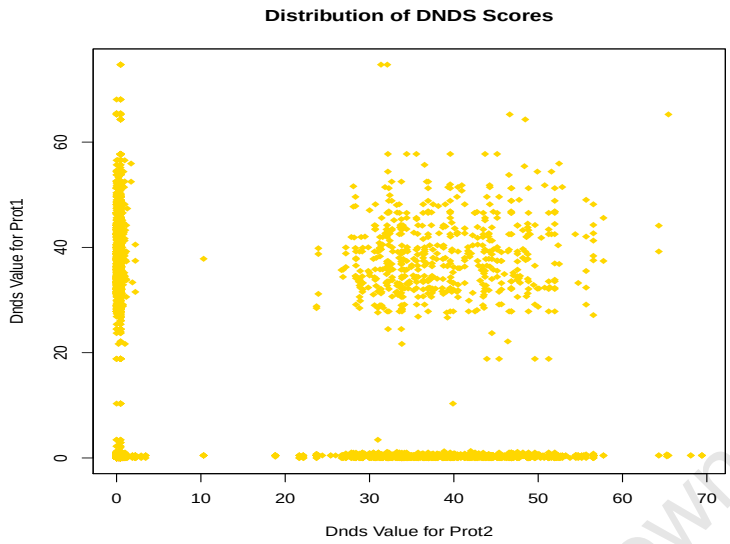


Figure 4.12: Distribution of dN/dS values obtained from the whole interaction set.

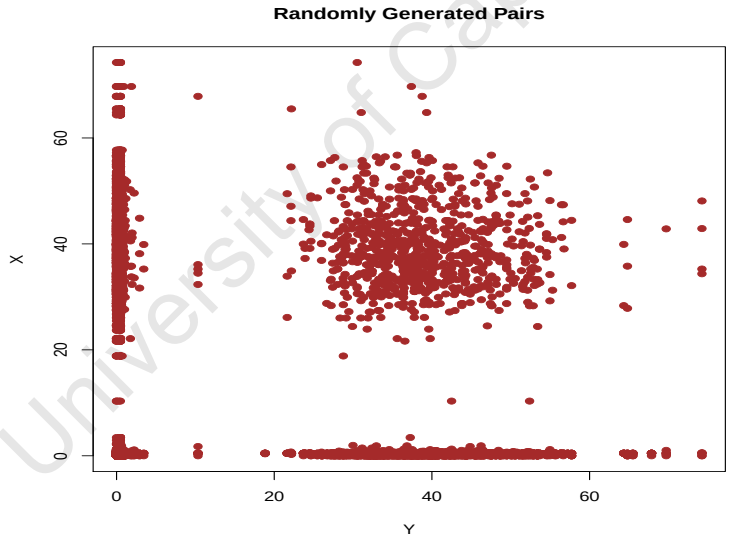


Figure 4.13: Distribution of randomly selected pairs of dN/dS values of the entire protein set.



### 4.5.1 Distribution of functional similarity scores

In Section 3.7, we hypothesized that interacting proteins perform similar functions, are involved in similar biological processes and are co-located. This hypothesis follows from

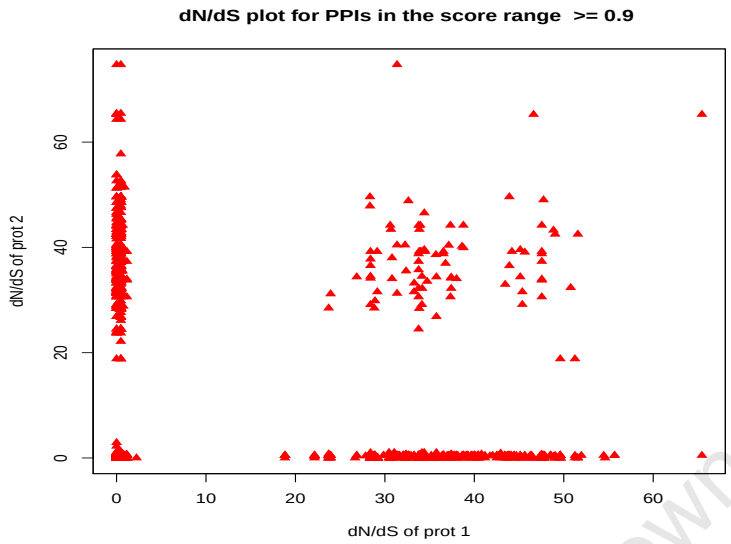


Figure 4.14: dN/dS plot of high confidence scores. This plot is similar to the one in Figure 4.12 above, again showing no real evidence of correlation in the dN/dS values for interacting proteins

the space-time constraints that must be in place to allow for a protein-protein interaction between a given pair of proteins.

We present in Figure 4.15, the the distribution of pairwise similarity scores for interacting proteins in the three different ontologies. From the graphs (histogram and density plots), it is evident that most pairwise similarity scores in the biological process ontology are close to 0, signifying low pairwise similarity. This observation can be explained by the fact that as at the time of the study, the whole-genome annotation of MTB proteins was still quite low, 52% according to Camus et al. (2002). Likewise for pairwise similarity for the PPIs in the molecular function ontology, most PPIs have low scores for pairwise similarity. However, in the plot for cellular component ontology, we observe a different distribution with peaks occurring both at the low score end, and the high score end. This observation can be attributed partly to the fact that we integrated different methods to predict subcellular localization hence providing a wider and more comprehensive coverage (See 3.2.2), and also to the requirement that proteins be present in the same cellular compartment in order to interact. The overall distribution of pairwise similarity across the ontologies is also summarised in the boxplot shown in Figure 4.16, which shows the distribution of pairwise similarity values across the quartiles. The observation that is clearly evident is the high pairwise similarity scores in the cellular component ontology across the whole network of predicted PPIs. The biological process ontology and the molecular function ontology both have a comparably similar distribution with the majority of the PPI pairwise similarity scores around 0.6.

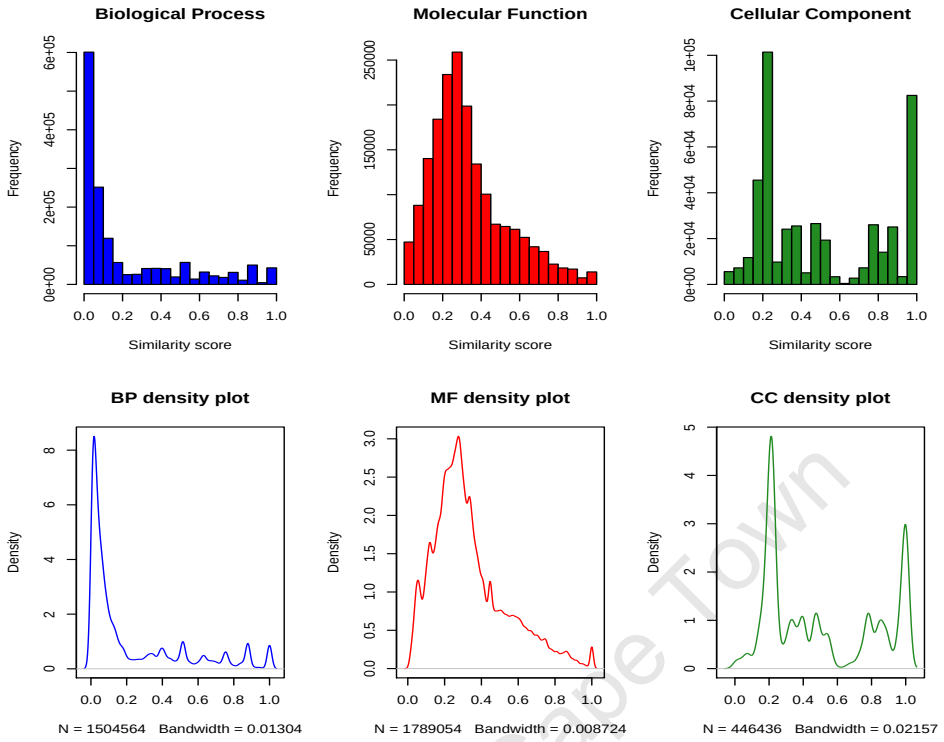


Figure 4.15: The figure above shows the distribution of similarity scores between protein pairs for the proteins that have GO annotations in the three ontologies respectively. Below each histogram is a density plot of the distribution offering a different perspective for viewing the data.

Figures 4.17 and 4.18 show the distribution taken for different total confidence score values. We noticed a shift of the peaks of the density plot curves from the low score end to the high score end in the molecular function ontology and the biological process ontology. This suggests that proteins constituting the high scoring PPI pairs share a lot in common with each other as opposed to the proteins in the low scoring region. In the cellular component pairwise similarity distribution, we observe a consistent trend of high functional similarity across the different score ranges.

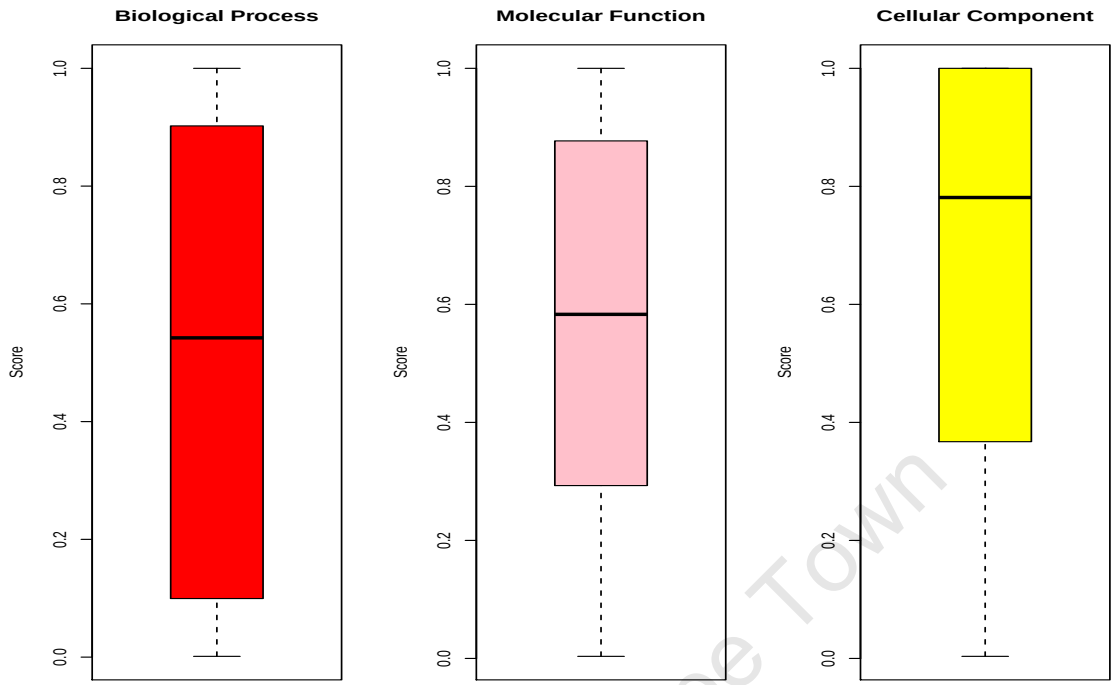


Figure 4.16: The graph above shows the distribution of GO similarity across the three ontologies. In all the cases, the median similarity measure is above 0.5 on a scale of 0.0-1.0, from low similarity to high similarity between the pairs, suggesting that more that half the PPIs in each ontology have both of the proteins constituting the pair sharing similar GO terms. In fact, the cellular component ontology displays the highest similarity values for proteins in PPI pairs where the median similarity is about 0.8.



### 4.5.2 Gene set enrichment analysis of top scoring protein-protein interactions

In this section, we look at the distribution of GO terms by depth in the GO DAG hierarchy. From this we decided to use the GO slim for the analysis, and we display graphically, the distribution of GO annotation across the different ontologies (BP, MF, CC). We then perform GO enrichment analysis to identify terms significantly over-or under-represented in the interaction set.

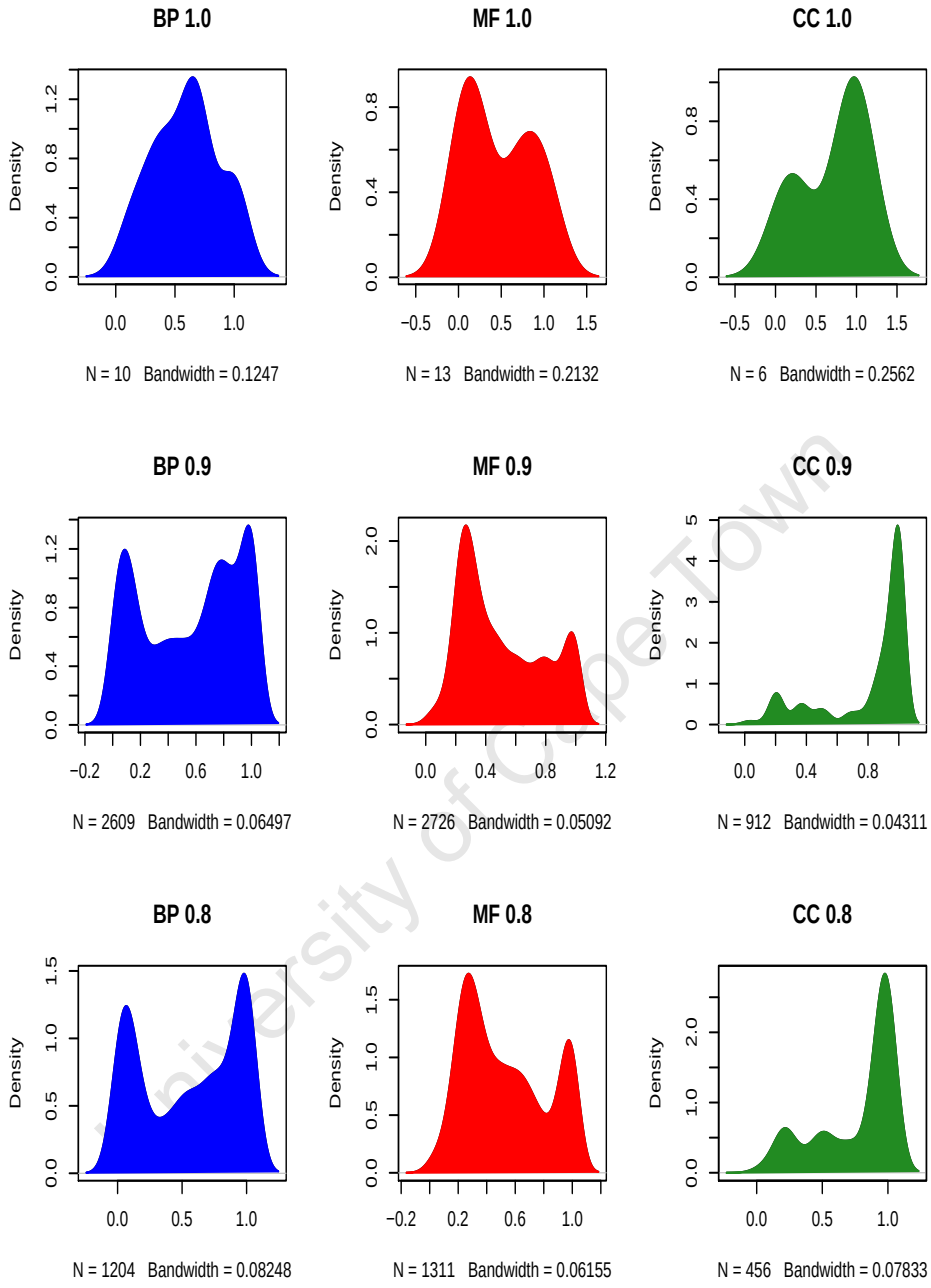


Figure 4.17: Density plots showing the distribution of similarity scores in the different ontologies for total scores greater than or equal to 0.8

### 4.5.3 GO level annotation

Figure 4.19 below shows the distribution of GO-levels for annotation of MTB proteins. The plot suggests that most proteins in the interaction network are annotated to general GO

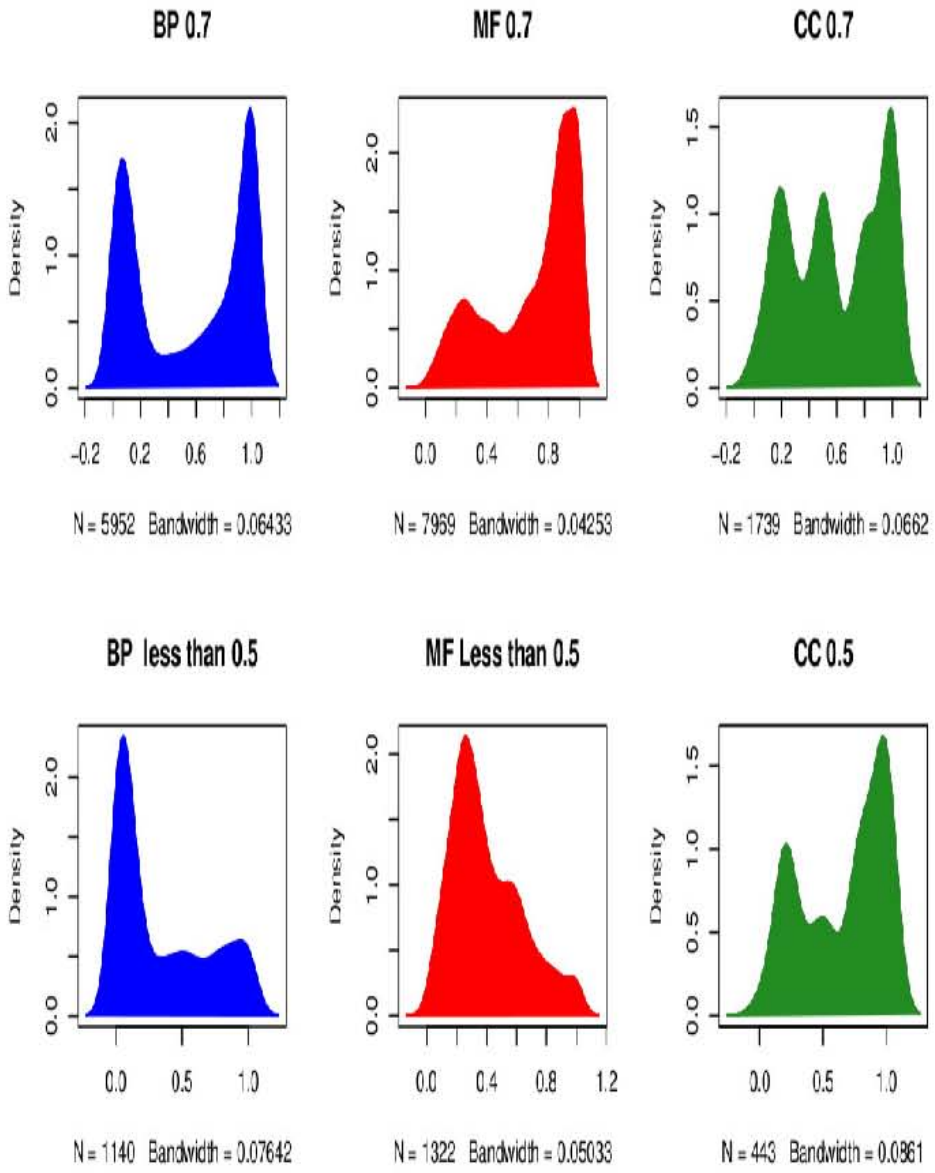


Figure 4.18: Density plots showing the distribution of similarity scores in the different ontologies for total score values less than 0.8

terms (GO-level values close to the roots of the ontologies). In fact looking at the graph, the levels that are highly represented are less than 5, except for the cellular component

ontology which has a significantly high number of proteins annotated to GO terms at level 5 or greater. GO level plots give an indication as to the depth of annotation of a particular genome. Due to the range in annotated levels we decided to use GO slim for the enrichment analysis.

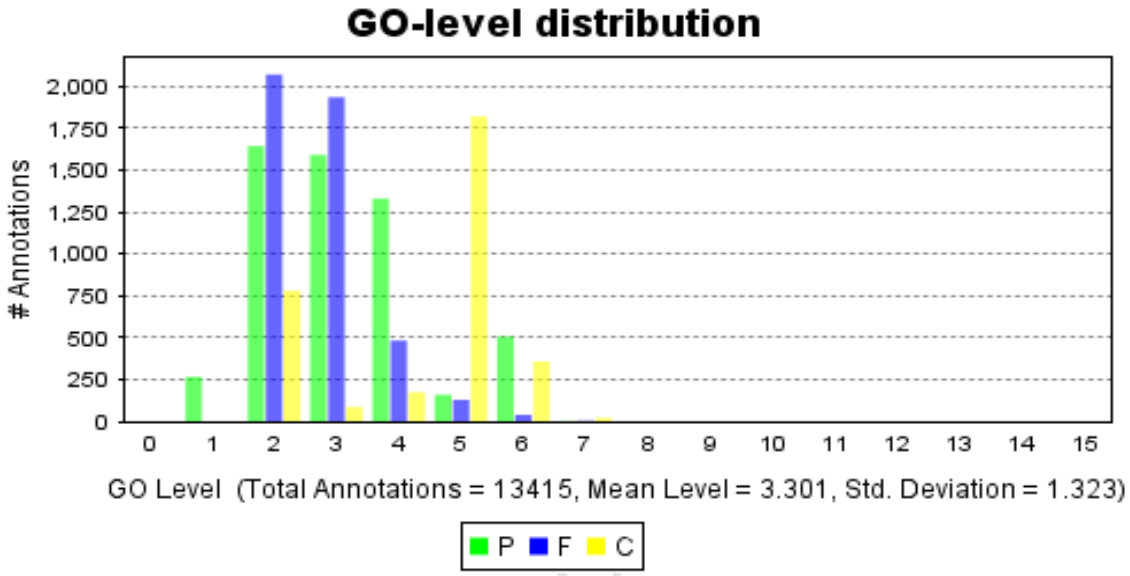


Figure 4.19: The graph above shows the distribution of GO annotation separated for each ontology–the different colours plus one letter abbreviation represent the three ontologies i.e , P represents Biological Process ontology, M, represents Molecular Function ontology and lastly C represents Cellular Component ontology.

### 4.5.4 GO distribution

We performed a GO term distribution and gene set enrichment analysis of proteins in PPIs with total scores in the range between 0.9 and 1.0, the highest possible score after the GO slimming process. The analysis is based on the Blast2GO Conesa et al. (2005), Conesa and Götz (2008) functional annotation tool. What we are trying to determine is whether some biological processes or molecular functions are particularly over-represented. These, as explained in the previous chapter could give leads as to the functions of proteins predicted to interact. In addition, we are also interested in the location of the interacting proteins, therefore we aimed to identify over-represented cellular components.

Firstly we represent the simple distribution of GO terms in each ontology. Figures 4.20

and 4.21, display the distribution of biological processes and molecular function terms respectively. Figure 4.20, shows that *transcription*, *biological process* and *response to stress* are well represented. Looking at Figure 4.22, it is evident that most of the proteins in the set examined are located in the cytoplasm and within the cell membrane.

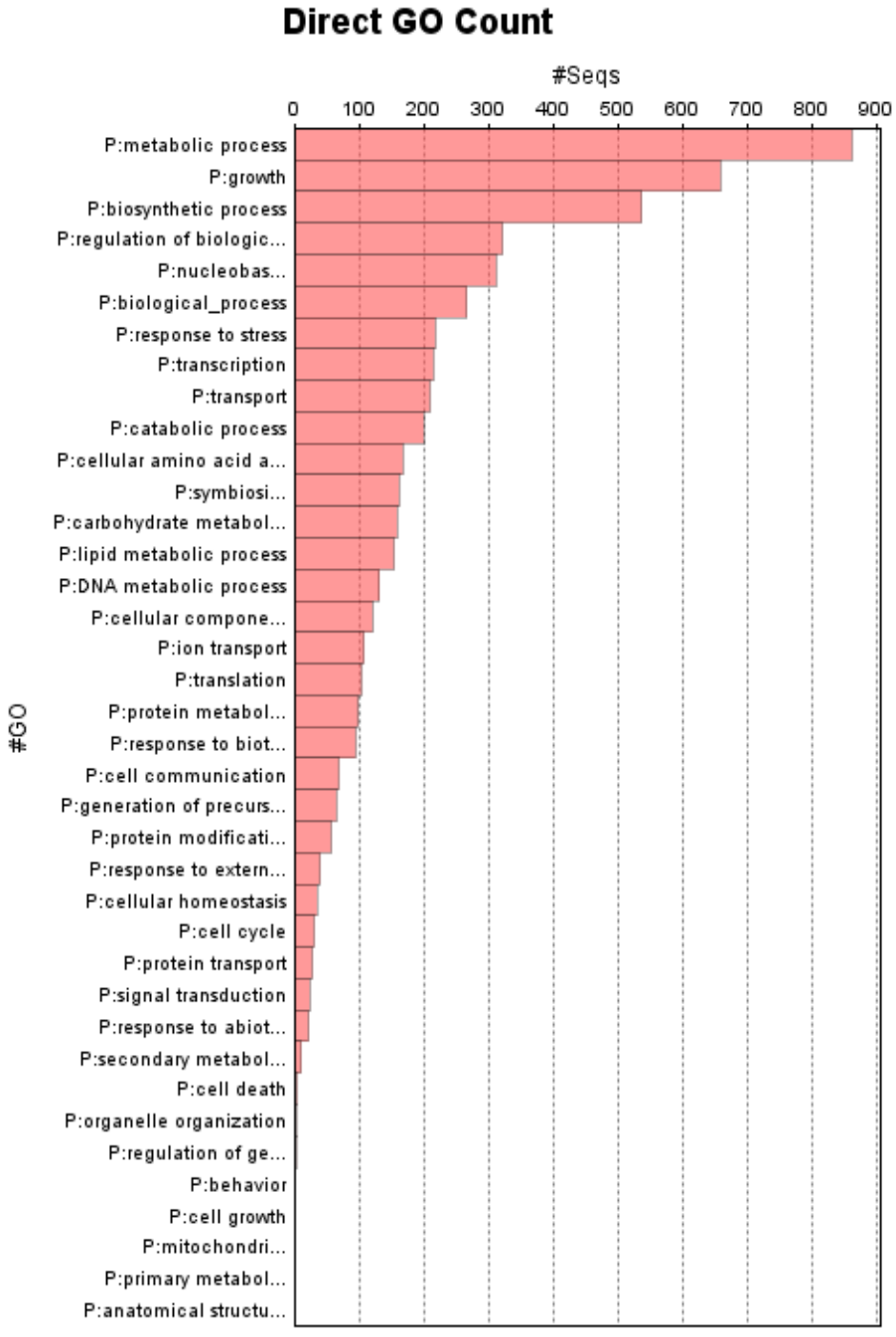


Figure 4.20: The graph above shows the direct GO term count of the slimmed Biological Process ontology.

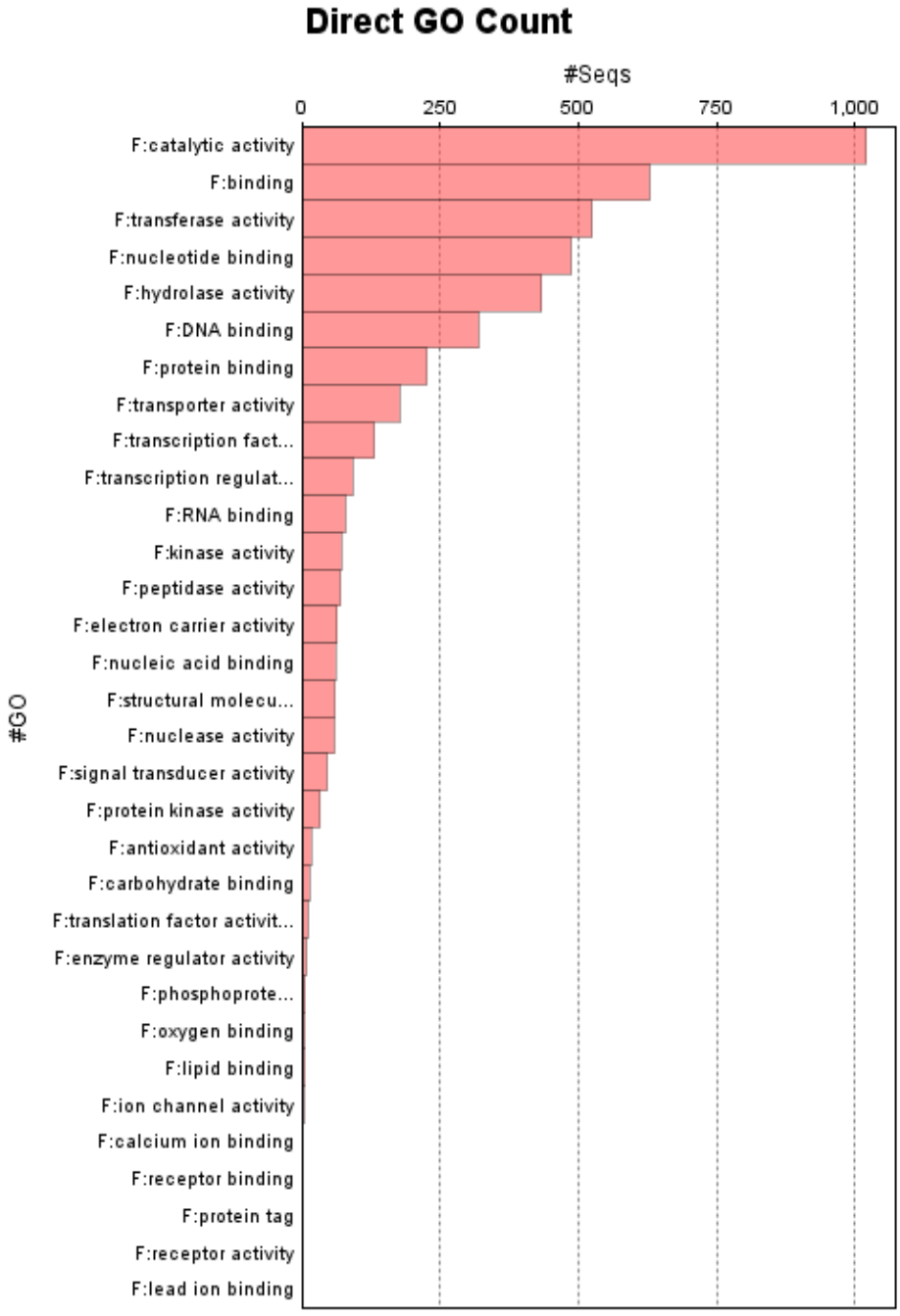


Figure 4.21: The graph above shows the direct GO term count of the slimmed Molecular Function ontology.

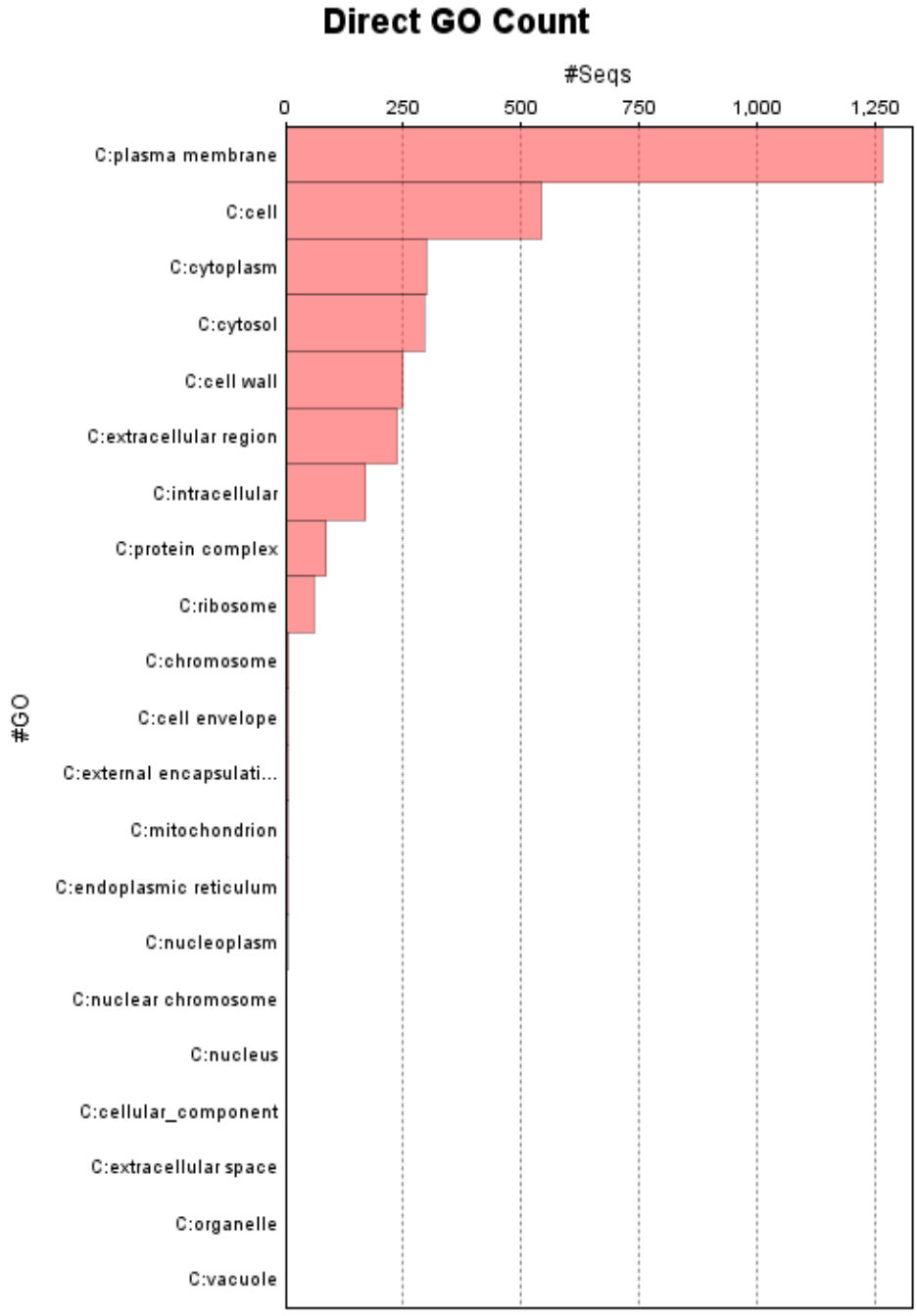


Figure 4.22: The graph above shows the direct GO term count of the slimmed Cellular Component ontology.

We performed GO enrichment analysis using GO slim terms for each ontology and the Fisher's exact test in BLAST2GO. The results for over-represented terms are shown in Table 4.6 and 4.7 at the end of this chapter. Table 4.5, in the next page shows terms that

are under-represented. Some of the interesting over-represented terms include different types of metabolic processes (this may be due to the use of KEGG pathways in STRING as a validator of interactions), protein complex, protein binding, and signal transduction. Interestingly, the under-represented terms were mostly related to transcription.

Table 4.5: Under-represented GO Terms.

| **GO ID** | **GO name** | **Ontology** | **P-Value** | **FDR** | **FWER** |
|---|---|---|---|---|---|
| GO:0050789 | regulation of biological process | process | 7.49E-11 | 0.00226781 | 0.00338751 |
| GO:0065007 | biological regulation | process | 2.75E-10 | 0.00226781 | 0.00338751 |
| GO:0003700 | transcription factor activity | function | 4.59E-08 | 0.00226781 | 0.003388 |
| GO:0003677 | DNA binding | function | 4.80E-07 | 0.00226781 | 0.00339321 |
| GO:0006350 | transcription | process | 7.94E-07 | 0.00226781 | 0.00339708 |
| GO:0030528 | transcription regulator activity | function | 1.19E-06 | 0.00226781 | 0.00340166 |

### 4.5.5 A brief analysis of some example predicted interactions

We discuss a few examples of predicted interactions to look at their biological plausibility. Most of the top hits with a confidence score of 1.0 and supported by 2 or more different evidence types (not including experiment), were interactions between 2 or more members of a protein complex. Protein P71811, the carbamoyl-phosphate synthase small chain protein, was predicted by STRING, DDI and the Ortholog prediction methods (total score of 1.0) to interact with P57689, the carbamoyl-phosphate synthase large chain. These proteins are annotated in UniProt to form a complex to catalyze the first step in L-arginine biosynthesis and pyrimidine metabolism, thus lending weight to our high confidence prediction. Another protein that STRING predicts P71811 to interact with at very high confidence, is P65613, aspartate carbamoyltransferase, which was identified in UniProt as a high-confidence drug target. Another example of a complex predicted by STRING, DDI and the Ortholog prediction methods (total score of 1.0) was the interaction between P94984, the phenylalanyl-tRNA synthetase alpha chain and P94985, the phenylalanyl-tRNA synthetase beta chain. Both were predicted to be located in the cytoplasm. A final example of a potential complex we correctly predicted (DDI method, and supported by STRING) is that of the interactions between P63852, the probable cytochrome c oxidase subunit 1, P63854, the cytochrome c oxidase subunit 2 and P63856, the probable cytochrome c oxidase subunit 3. All of these are also predicted be co-located in the cell membrane, and P63854 was identified in UniProt as a high-confidence drug target. There are many more examples of protein complex interactions that we correctly predicted using the DDI or ortholog prediction approaches.



We also noted that transcription is underrepresented in the subset of MTB PPIs that we identified despite transcription regulation network being identified by Sanz et al. (2011) as one of the most prominent regulatory networks in MTB. This can be investigated further as a future extension of the project.

One of the interactions predicted by DDI and supported by STRING and experimental evidence was the interaction between Q11034, an uncharacterized protein predicted to be involved in transcription regulation, and Q11035, an uncharacterized protein predicted to be an anti-sigma factor, and thus would be involved in transcription regulation. Anti-sigma factors bind to sigma factors to affect transcription regulation. A second interaction predicted by DDI and supported by STRING and experimental evidence is that

between Q11053, a probable serine/threonine-protein kinase pknH and P66799, a probable regulatory protein embR. These proteins probably interact in a two-component signal transduction system. We also predicted an interaction at high confidence (score of 1.0) between P0A5L0, the peptide methionine sulfoxide reductase msrA, which is an important repair enzyme in oxidative stress, and P71971, the PilB-related protein uncharacterized protein, whose GO annotation suggests involvement in oxidation-reduction and response to oxidative stress. This may help us to characterise such putative proteins.

To provide some interaction examples that we predicted that were not predicted by STRING, a combination of DDI using IntAct data and Orthologs predicted a high confidence interaction between P66753, the Holliday junction ATP-dependent DNA helicase ruvB, involved in DNA damage repair and the stress response, and P66028, transcription termination factor Rho, which was identified as a high-confidence drug target. P66753 was also predicted, through ortholog evidence only (score of 0.75), to interact with P0A5B9, the chaperone protein dnaK and P0A548, the chaperone protein dnaJ, both of which are involved in heat shock and stress response. The latter was predicted to be a high confidence drug target. STRING also correctly predicts the known interaction between RuvB (P66753 and RuvA (P66744). In a final example, DDI using IntAct data and Orthologs predicted a high confidence interaction between P68909, uncharacterized protein Rv2897c/MT2965, identified as a high-confidence drug target with peptidase activity (GO annotation) and P66842, a cell division protein ftsY homolog, which may be involved in targeting proteins to the membrane. Again, this interaction may help us to better understand the function of the currently uncharacterised protein. While we can not describe all interactions predicted, the examples chosen provide some biological plausibility to the predictions and highlight some potentially interesting interactions with predicted drug targets.

Table 4.6: Over-represented GO terms

| GO ID | GO name | Ontology | P-Value | FDR | FWER |
|---|---|---|---|---|---|
| GO:0006091 | generation of precursor metabolites and energy | process | 0 | 1.80E-10 | 1.18E-09 |
| GO:0043170 | macromolecule metabolic process | process | 0 | 1.80E-10 | 1.18E-09 |
| GO:0008152 | metabolic process | process | 0 | 1.80E-10 | 1.18E-09 |
| GO:0009058 | biosynthetic process | process | 0 | 1.80E-10 | 1.18E-09 |
| GO:0044238 | primary metabolic process | process | 1.92E-13 | 1.80E-10 | 1.18E-09 |
| GO:0044237 | cellular metabolic process | process | 1.99E-13 | 1.80E-10 | 1.18E-09 |
| GO:0044267 | cellular protein metabolic process | process | 4.48E-13 | 1.80E-10 | 1.18E-09 |
| GO:0044260 | cellular macromolecule metabolic process | process | 4.48E-13 | 1.80E-10 | 1.18E-09 |
| GO:0009056 | catabolic process | process | 4.69E-13 | 1.80E-10 | 1.18E-09 |
| GO:0009059 | macromolecule biosynthetic process | process | 5.21E-13 | 1.80E-10 | 1.18E-09 |
| GO:0044249 | cellular biosynthetic process | process | 5.21E-13 | 1.80E-10 | 1.18E-09 |
| GO:0006412 | translation | process | 5.21E-13 | 1.80E-10 | 1.18E-09 |
| GO:0006519 | amino acid and derivative metabolic process | process | 7.68E-13 | 1.80E-10 | 1.20E-09 |
| GO:0005737 | cytoplasm | component | 7.82E-13 | 1.80E-10 | 1.20E-09 |
| GO:0019538 | protein metabolic process | process | 1.07E-12 | 1.80E-10 | 1.21E-09 |
| GO:0003723 | RNA binding | function | 1.12E-12 | 1.80E-10 | 1.21E-09 |
| GO:0000166 | nucleotide binding | function | 1.42E-12 | 1.80E-10 | 1.21E-09 |
| GO:0040007 | growth | process | 1.47E-12 | 1.80E-10 | 1.21E-09 |
| GO:0005488 | binding | function | 1.49E-12 | 1.80E-10 | 1.21E-09 |
| GO:0005622 | intracellular | component | 1.65E-12 | 1.80E-10 | 1.21E-09 |
| GO:0032991 | macromolecular complex | component | 1.71E-12 | 1.80E-10 | 1.21E-09 |
| GO:0003824 | catalytic activity | function | 1.84E-12 | 1.80E-10 | 1.21E-09 |
| GO:0009987 | cellular process | process | 1.85E-12 | 1.80E-10 | 1.21E-09 |
| GO:0044424 | intracellular part | component | 2.30E-12 | 1.80E-10 | 1.21E-09 |
| GO:0016740 | transferase activity | function | 2.68E-12 | 1.80E-10 | 1.21E-09 |
| GO:0043234 | protein complex | component | 3.49E-12 | 1.80E-10 | 1.21E-09 |
| GO:0043232 | intracellular non-membrane-bound organelle | component | 9.65E-12 | 1.83E-10 | 1.35E-09 |

Table 4.7: Over-represented GO terms (cont.)

| GO ID | GO name | Ontology | P-Value | FDR | FWER |
|---|---|---|---|---|---|
| GO:0043228 | non-membrane-bound organelle | component | 9.65E-12 | 1.83E-10 | 1.35E-09 |
| GO:0005975 | carbohydrate metabolic process | process | 1.62E-11 | 1.83E-10 | 1.37E-09 |
| GO:0005840 | ribosome | component | 2.39E-10 | 4.75E-10 | 3.80E-09 |
| GO:0030529 | ribonucleoprotein complex | component | 2.39E-10 | 4.75E-10 | 3.80E-09 |
| GO:0005198 | structural molecule activity | function | 4.77E-10 | 8.53E-10 | 7.04E-09 |
| GO:0043229 | intracellular organelle | component | 5.66E-10 | 9.82E-10 | 8.35E-09 |
| GO:0044444 | cytoplasmic part | component | 3.27E-07 | 4.67E-07 | 4.32E-06 |
| GO:0016020 | membrane | component | 5.21E-07 | 7.14E-07 | 6.96E-06 |
| GO:0005886 | plasma membrane | component | 5.21E-07 | 7.14E-07 | 6.96E-06 |
| GO:0016772 | transferase activity, transferring phosphorus-containing groups | function | 1.72E-06 | 2.33E-06 | 2.39E-05 |
| GO:0016301 | kinase activity | function | 1.72E-06 | 2.33E-06 | 2.39E-05 |
| GO:0016043 | cellular component organization and biogenesis | process | 2.35E-06 | 3.34E-06 | 3.51E-05 |
| GO:0006139 | nucleobase, nucleoside, nucleotide and nucleic acid metabolic process | process | 7.14E-06 | 9.62E-06 | 1.03E-04 |
| GO:0005515 | protein binding | function | 4.28E-05 | 5.78E-05 | 6.36E-04 |
| GO:0030312 | external encapsulating structure | component | 2.01E-04 | 2.81E-04 | 0.003157 |
| GO:0005618 | cell wall | component | 2.30E-04 | 3.07E-04 | 0.00351889 |
| GO:0016787 | hydrolase activity | function | 3.48E-04 | 4.60E-04 | 0.00539309 |
| GO:0005829 | cytosol | component | 5.55E-04 | 7.16E-04 | 0.00854995 |
| GO:0005215 | transporter activity | function | 6.22E-04 | 8.03E-04 | 0.00978389 |
| GO:0043283 | biopolymer metabolic process | process | 6.50E-04 | 8.29E-04 | 0.0103147 |
| GO:0007165 | signal transduction | process | 7.11E-04 | 9.14E-04 | 0.0115863 |
| GO:0008135 | translation factor activity, nucleic acid binding | function | 8.89E-04 | 0.00114644 | 0.0150769 |
| GO:0045182 | translation regulator activity | function | 8.89E-04 | 0.00114644 | 0.0150769 |

# Chapter 5

# Discussion



This chapter discusses the significance of the results that we obtain from this study in light of the objectives set out in Chapter 1, which include: to predict protein-protein interactions computationally; to evaluate molecular evolution dynamics of the predicted protein-protein interactions, and to biologically interpret the predicted protein-protein interactions.

The main objectives of this study were first to implement and evaluate a select set of methods that are currently used to predict PPIs, and secondly, to describe two algorithms, Ortholog Prediction of Interaction Algorithm (OPIA), and Domain Evidence Algorithm (DEA) which use the concepts of orthologous interaction transfer Walhout et al. (2000), Yellaboina et al. (2008), and domain-domain interactions Ng et al. (2003b), to infer protein-protein interactions respectively (see Chapter 2). We also implemented an integrative scoring scheme in Section 3.4.2 to compute total confidence scores for PPIs predicted by disparate methods. In addition, we integrated subcelluar localization information which we get from data mining the UniProt database, and from running Psort 3.0 on the entire protein set of the MTB genome. We then performed functional analysis on proteins in our PPI set. We performed this analysis, by first running the Blast2GO program Conesa et al. (2005), and finding functional categories that are overrepresented with the interest of determining the functions of the proteins that we predict to interact. In the following paragraphs, I'll discuss in-depth the findings of this study.

The main reason for doing computational predictions is the small number of known PPIs in the database of MTB. Our motivation for integrating different methods to infer PPIs

follow the flaws found in the various methods that have conventionally been used to infer PPIs including the ones that are highlighted in Section 2.2.1.

In the Ortholog Prediction of Interactions Algorithm (OPIA), we infer interactions in MTB from known orthologs in other organisms. Orthologous proteins have in some cases been known to maintain the same functional interactions making possible transfer of functions to related proteins in other species Yu et al. (2004). Protein interactions have also been observed to be conserved across species Sharan et al. (2005). As at the end of December 2010, there were 271,764 unique interactions in the IntAct database Aranda et al. (2010), our data source for interactions inferred by OPIA. We were able to identify 1702 interactions in MTB by the use of OPIA (see details in the method section in Chapter 2). It should be however be noted that this method is also not free from false positives. These false positives are attributed to the inherent flaws (see Section 2.2.4) in the various methods that are used to predict the PPIs that are stored in the IntAct database, in addition to the fact that in some cases protein interactions may not be transferable, especially in cases where the reference organisms are phylogenetically distant Yellaboina et al. (2008). These possible flaws are however partly handled in the scoring scheme, in which the methods that are used to infer the interactions are weighted differently depending on the confidence that we attach to them. In addition, we integrate all the contributing evidence scores for PPIs that are supported by more than one evidence. Scores for PPIs predicted by interologs are set at 0.75.

Protein interaction networks can be decomposed into their constituent domain interaction networks for the purposes of elucidating both structurally and functionally, the finer details of DDIs Albrecht et al. (2005), Pagel et al. (2008), Raghavachari et al. (2008), Deng et al. (2002). The Domain Evidence Algorithm (DEA) uses the idea that PPIs are DDIs taken at a more general level Albrecht et al. (2005). We found protein domains that have been observed to interact in other organisms and suggested that they possibly interact in MTB. The next process involved the retrieval of all MTB proteins containing these predicted interacting domains and inferred that they interact too. We categorized domain interactions into two groups, first are the DDIs deemed to be of high quality as a result of being supported by their 3-D structure evidence from the PDB database Sussman et al. (1998). The second category is that of domain interactions whose 3-D structure evidence from PDB was not determined at the time of the study. We set a score of 0.75 to DDI with 3-D evidence and a score of 0.7 to DDI without 3-D structure evidence.

We also obtained functional interactions of MTB proteins from the STRING database. The STRING database von Mering et al. (2007), as explained in the methods section in Chapter 3, integrates different methods to infer functional interactions. To score these PPIs, we used the raw scores obtained from STRING (see 3.2.1 for score computation). We obtained 20261 PPIs from STRING, which are proportionally the largest contributant to the total number of unique interactions (27569) that we identified in this study. The PPI scores range between (0-1.0), 1.0 being the highest possible score depicting a 'true' interaction.

The last category of PPIs, a total of 53 interactions were obtained from experiments in the IntAct database. These experimental scores are considered to be of the highest quality, of all the PPIs therefore we attach a relative score value of 1.0 to these.

Following the distribution of interactions outlined above, we have observed that the majority of interactions inferred in this study are from interactions derived from the STRING database, with those obtained by domain interactions coming a distant second. An explanation is the fact that STRING integrates protein interaction data from many different sources hence providing a wide, and therefore more comprehensive coverage of PPIs. In addition, STRING aims to generate not only physical protein-protein interactions but also other functional interactions. A method like OPIA, which relies on interaction transfer across orthologs is greatly limited by how well a protein in a given proteome has been studied plus the number of its orthologs available in other species. In this case, not many MTB proteins have known functions, genome annotation of MTB proteins stand at about 52 % Camus et al. (2002). When not much is known about two proteins predicted to interact, it is obviously difficult to gauge the plausibility of such an interaction by considering a spatial factor such as the subcellular localization, which would give leads as to whether the localization of the proteins within the cell support the possibility of their interaction. It would also be of interest to know whether the two proteins predicted to interact are expressed at the same time in order to account for the time factor needed for a PPI to take place.

A density plot in Figure 4.8 showed that most of the dN/dS values that were calculated for MTB proteins in this research clustered around values less than 1. Sequence divergence has been observed to be related to protein dispensability, due to strong purifying selection on less dispensable proteins. Dispensability of a protein to a large extent depends on how

central a protein is, that is how many other proteins, a given protein is connected to either physically or functionally. Teichmann (2002), discovered that proteins that form stable complexes (therefore considered central) are more evolutionarily conserved than those that form transient complexes or those that have not been observed to participate in protein-protein interactions, which were observed to be the least conserved of the lot. A study by Valdar and Thornton (2001), in which they compared sequence conservation among six homodimers, revealed that amino acid residues at protein interfaces are relatively more conserved than their counterparts in other regions of the protein sequence. These observations add weight evidence to the use of orthologs to predict protein-protein interactions in the organism of tuberculosis, as they support the notion that orthologs interact in a similar way given the conservation of interaction interfaces.

We examined GO enrichment of the proteins in our interactions set and measured GO similarity scores between interacting proteins. We generated histograms and density plots of total similarity scores of all the interacting pairs in the three ontologies (biological process, molecular function and cellular component) (Figure 4.15). We performed density plots of interactions in bands segmented by PPI confidence score ranges. The density plots are done separately for the three ontologies. The observation here is that, unlike in the density plots for the entire interaction network that we predicted (see Figure 4.15), where few proteins fall under high, (greater than 0.7) interaction scores, we see that the distribution of PPIs skew towards higher GO similarity scores progressively as the total confidence score increases with some exceptions, as shown in Figure 4.17. Surprisingly, the molecular function scores seemed to perform better in the 0.7 confidence range. For BP and CC, however, the higher confidence interactions tended to have higher similarity scores. This observation suggests a general correlation between interaction scores and the similarity scores across the ontologies in the protein pairs considered. Following this realization we selected PPIs whose confidence scores fall between 0.9 and 1.0 for GO term enrichment investigation.

We analyze the GO slim terms (see Figure 3.8 for GO-Slimming process) of interactions obtained in the previous paragraph. A closer look at Figure 4.19 which shows the GO level distribution per ontology shows that the mean depth level in the GO hierarchy, representing how specific the GO-terms are is about 3. A possible explanation for this observation is that most MTB proteins have not been clearly annotated hence most terms just represent general annotations–terms that are closer to the root of the ontologies Fleischmann et al.

(2002). The distribution of GO terms associated with proteins in the interaction confidence score bracket of 0.9-1.0, across three ontologies BP, MF, and CC are shown in the following graphs Figure 4.20 , 4.21, 4.22 respectively.

The graph for biological process shows that *response to stress* and *metabolic process* are some of the biological processes with high direct GO counts. Bacterial pathogens are known to survive under two entirely different environments, namely, their natural habitat and that of the cells of their host Chowdhury et al. (1996). The virulence determinants of pathogenic bacteria are under the control of transcription activators which respond to stress caused by factors such as temperature, osmolarity, metal ion concentration and oxygen tension of the environment Chowdhury et al. (1996). Proteins involved in response to stress–more specifically for this study, the protein-protein interactions that occur during stress response would be key to understanding the virulence of MTB.

In a close analysis of the molecular function ontology graph in Figure 4.21, we found for example that functions associated with *binding* are particularly elevated. This is an interesting revelation as this provides additional validation that these proteins are highly likely to interact following their shared functional class that supports protein-protein interactions.

The distribution of GO terms in the cellular component (CC) ontology skews towards membrane proteins and cytoplasmic proteins. It would be interesting to further examine the interaction partners of membrane proteins in this category when studying host-pathogen PPIs, which is beyond the scope of this study. Membrane proteins and secreted proteins are more likely to be the interacting partners with the host proteins in comparison to, for example, cellular proteins.

We performed GO enrichment analysis on high confidence PPIs from our set of predicted interactions. Tables 4.6 and 4.7 show GO terms that are over-represented in our set of PPIs as compared to the whole genome. Some of the over-represented biological processes are *metabolic process* and *signal transduction*. Living cells require cell metabolism in order to grow, reproduce, maintain their structures and perceive stimuli from their living environment. Studying PPIs that occur in order to achieve the functions mentioned above is important in understanding the survival mechanism of an organism.

Signal transduction is the process by which stimuli from the environment external to the cell are relayed to the cell through the cell membrane in order to activate intracellular responses

King (2011). Signal transduction process is a receptor-specific process Prahlad T. Ram, Ravi Iyengar (2008), in the sense that a given receptor will only activate a specific set of signaling components downstream thereby aiding PPIs in the signaling process at different stages to be studied in isolation. Proteins that interact to perform signal transduction are key in understanding the invasion mechanism of the pathogen and the host immune response activated in the cell during the invasion.

In molecular function ontology, *binding*, *protein binding* and *catalytic activity* are over-represented. The over-representation of metabolic processes supports the over-representation of binding activity since proteins bind to one another in order to carry out the metabolic processes. Catalytic activity is also part of the metabolic process.

Looking at the cellular component ontology, *cytoplasm*, *membrane*, *macromolecular complex* are over-represented. Membranic and cytoplasmic proteins are important for both metabolic processes and signal transduction, as explained above.

# Chapter 6

# Concluding Remarks



Experimental protein interaction studies in MTB are limited by factors such as length of time and costly experimental design and setup. We proposed and implemented two computational algorithms for inferring interacting proteins OPIA and DEA. We have demonstrated that different data sources can be integrated together to predict the possible occurrence of PPIs.

We have developed a scoring scheme, that scores the confidence with which we view a particular protein-protein interaction. The scoring scheme takes into account the number of independent evidences supporting a particular interaction. We have shown that interactions that are supported by multiple independent evidences, score higher than interactions supported by few evidences. A possible improvement on the prediction methods that we have proposed in this study, would be to integrate more data from different sources including additional prediction methods, and data mining the literature to provide more evidence supporting our predicted interactions. Just like with any other computational prediction, the interactions that we have predicted in this study are still subject to experimental tests to ascertain their validity.

Another important aspect of PPIs that would further increase the confidence that we attach to the PPIs that we identified using this method, and in general other computational methods used to predict PPIs, is to integrate gene expression data that would provide spatio-temporal data. Spatio-temporal data is needed to validate interaction in time and space since structural evidence alone is not enough to ascertain true PPIs.

We also evaluated the evolutionary dynamics of the MTB proteins and found that most MTB proteins are under purifying selection, which alludes to the findings of previous studies, where there is an inverse relationship between sequence divergence and protein dispensability Teichmann (2002). In this case, a protein's dispensability is quantified by the number of interaction partners that a protein has.

We have also investigated the GO terms that are particularly enriched for high scoring interactions. We have identified that GO terms associated with biological processes such as *response to stress* which are overrepresented. It would be interesting for further studies to investigate the possible cause of this observation and whether proteins involved in these processes could be possible drug targets.

A new method for scoring functional similarity has also been used in this study. The method uses only the topology of the GO DAG to infer similarity unlike the current methods that use information content (IC). This is a big step forward towards mitigating functional similarity calculation challenges, which include inaccuracy in calculating similarity by IC based methods due to the ever changing GO database.

All in all, computational methods provide relatively less expensive ways of inferring PPIs, however the truth of these interactions are only convincingly validated when these predicted PPIs are subjected to experimental tests, possibly *in vivo*.

This work can further be extended by (1) incorporating additional data sources including protein interaction data from text mining literature databases for PPI data (2) developing a webservice that retrieves data in real time from the databases used instead of working with downloaded flat files to ensure comprehensive and the most up-to-date information coverage.

Another avenue that this work can be extended to is host-pathogen interaction PPI study carried out between human proteins and those of MTB. This is possible due to the fact that most of the methods discussed here are applicable accross species.

The scoring scheme used for estimating the interaction confidence can also be improved upon to allow for more robustness instead of relying heavily on the theoritical confidence attached to the detection methods. This is particularly important due to the fact that even experimentally derived PPIs have been shown to contain false positives. The confidence

that is attached to a particular interaction can also be optimized to include such qualities as the number of organisms that a particular PPI has been transferred as well as how close to MTB is the organism in which an interaction has been transferred. Currently all PPIs in MTB interologs are treated equally which is not ideal given that one can hypothesize that the evolutionary distance between two organisms plays a part in whether or not a PPI is preserved for a particular reason or is just a ubiqtuous occurance.

As an extension of this work, we plan to develop a database that would maintain the data generated in this study and deploy it on the internet for public access.

# Bibliography


Albrecht, M., Huthmacher, C., Tosatto, S. C. E., and Lengauer, T. (2005). Decomposing protein networks into domain-domain interactions. *Bioinformatics*, 21 Suppl 2:ii220–ii221. 28, 88

Aranda, B., Achuthan, P., Alam-Faruque, Y., Armean, I., Bridge, A., Derow, C., Feuermann, M., Ghanbarian, A. T., Kerrien, S., Khadake, J., Kerssemakers, J., Leroy, C., Menden, M., Michaut, M., Montecchi-Palazzi, L., Neuhauser, S. N., Orchard, S., Perreau, V., Roechert, B., van Eijk, K., and Hermjakob, H. (2010). The intact molecular interaction database in 2010. *Nucleic Acids Res*, 38(Database issue):D525–D531. 14, 34, 36, 88

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet*, 25(1):25–29. 41

Barrell, D., Dimmer, E., Huntley, R. P., Binns, D., O'Donovan, C., and Apweiler, R. (2009). The goa database in 2009–an integrated gene ontology annotation resource. *Nucleic Acids Res*, 37(Database issue):D396–D403. 41, 42

Ben-Hur, A. and Noble, W. S. (2005). Kernel methods for predicting protein-protein interactions. *Bioinformatics*, 21 Suppl 1:i38–i46. 15, 28

Bhardwaj, N. and Lu, H. (2005). Correlation between gene expression profiles and protein-protein interactions within and across genomes. *Bioinformatics*, 21(11):2730–2738. 19, 50

Biswas, M., O'Rourke, J. F., Camon, E., Fraser, G., Kanapin, A., Karavidopoulou, Y., Kersey, P., Kriventseva, E., Mittard, V., Mulder, N., Phan, I., Servant, F., and Apweiler, R. (2002). Applications of interpro in protein annotation and genome analysis. *Brief Bioinform*, 3(3):285–295. 43

Blaschke, C., Hirschman, L., and Valencia, A. (2002). Information extraction in molecular biology. *Brief Bioinform*, 3(2):154–165. 23

Bork, P. (1991). Shuffled domains in extracellular proteins. *FEBS Lett*, 286(1-2):47–54. 8

Braun, P., Tasan, M., Dreze, M., Barrios-Rodiles, M., Lemmens, I., Yu, H., Sahalie, J. M., Murray, R. R., Roncari, L., de Smet, A.-S., Venkatesan, K., Rual, J.-F., Vandenhaute, J., Cusick, M. E., Pawson, T., Hill, D. E., Tavernier, J., Wrana, J. L., Roth, F. P., and Vidal, M. (2009). An experimentally derived confidence score for binary protein-protein interactions. *Nat Methods*, 6(1):91–97. 14, 17

Brown, K. R. and Jurisica, I. (2005). Online predicted human interaction database. *Bioinformatics*, 21(9):2076–2082. 20

Burns, N., Grimwade, B., Ross-Macdonald, P. B., Choi, E. Y., Finberg, K., Roeder, G. S., and Snyder, M. (1994). Large-scale analysis of gene expression, protein localization, and gene disruption in saccharomyces cerevisiae. *Genes Dev*, 8(9):1087–1105. 10

Camus, J.-C., Pryor, M. J., Médigue, C., and Cole, S. T. (2002). Re-annotation of the genome sequence of mycobacterium tuberculosis h37rv. *Microbiology*, 148(Pt 10):2967–2973. 72, 89

Chalfie, M., Tu, Y., Euskirchen, G., Ward, W. W., and Prasher, D. C. (1994). Green fluorescent protein as a marker for gene expression. *Science*, 263(5148):802–805. 10

Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T. J., Higgins, D. G., and Thompson, J. D. (2003). Multiple sequence alignment with the clustal series of programs. *Nucleic Acids Res*, 31(13):3497–3500. 47

Chothia, C. and Janin, J. (1975). Principles of protein-protein recognition. *Nature*, 256(5520):705–708. 13

Chowdhury, R., Sahu, G. K., and Das, J. (1996). Stress response in pathogenic bacteria. *Journal of Bioscience*, 21:149–160. 91

Codoner, F. M. and Fares, M. A. (2008). Why should we care about molecular coevolution? *Evol Bioinform Online*, 4:29–38. 24, 25, 26, 48

Cole, S. T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S. V., Eiglmeier, K., Gas, S., Barry, C. E., 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, M. A., Rajandream, M. A., Rogers, J., Rutter, S., Seeger, K., Skelton, J., Squares, R., Squares, S., Sulston, J. E., Taylor, K., Whitehead, S., and Barrell, B. G. (1998). Deciphering the biology of mycobacterium tuberculosis from the complete genome sequence. *Nature*, 393(6685):537–544. 2

Conesa, A. and Götz, S. (2008). Blast2go: A comprehensive suite for functional analysis in plant genomics. *Int J Plant Genomics*, 2008:619832. 77

Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talán, M., and Robles, M. (2005). Blast2go: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, 21(18):3674–3676. 53, 77, 87

Cox, R. A. (2004). Quantitative relationships for specific growth rates and macromolecular compositions of mycobacterium tuberculosis, streptomyces coelicolor a3(2) and escherichia coli b/r: an integrative theoretical approach. *Microbiology*, 150(Pt 5):1413–1426. 2

Dandekar, T., Snel, B., Huynen, M., and Bork, P. (1998). Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem Sci*, 23(9):324–328. 19, 50

Deng, M., Mehta, S., Sun, F., and Chen, T. (2002). Interactions inferring domain â domain interactions from protein â protein interactions. *Genome Research*, pages 1540–1548. 29, 30, 88

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*, 95(25):14863–14868. 19

Emanuelsson, O., Brunak, S., von Heijne, G., and Nielsen, H. (2007). Locating proteins in the cell using targetp, signalp and related tools. *Nat Protoc*, 2(4):953–971. 9, 10

Enright, A. J., Iliopoulos, I., Kyrpides, N. C., and Ouzounis, C. A. (1999). Protein interaction maps for complete genomes based on gene fusion events. *Nature*, 402(6757):86–90. 17, 18, 29, 30

Fares, M. A. and McNally, D. (2006). Caps: coevolution analysis using protein sequences. *Bioinformatics*, 22(22):2821–2822. 24

Fields, S. and Johnston, M. (2005). Cell biology. whither model organism research? *Science*, 307(5717):1885–1886. 35

Finn, R. D., Marshall, M., and Bateman, A. (2005). ipfam: visualization of protein-protein interactions in pdb at domain and amino acid resolutions. *Bioinformatics*, 21(3):410–412. 30

Finn, R. D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J. E., Gavin, O. L., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E. L. L., Eddy, S. R., and Bateman, A. (2010). The pfam protein families database. *Nucleic Acids Res*, 38(Database issue):D211–D222. 28

Finn, R. D., Tate, J., Mistry, J., Coggill, P. C., Sammut, S. J., Hotz, H.-R., Ceric, G., Forslund, K., Eddy, S. R., Sonnhammer, E. L. L., and Bateman, A. (2008). The pfam protein families database. *Nucleic Acids Res*, 36(Database issue):D281–D288. 33

Fisher, R. A. (1922). On the interpretation of chi square from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, 85(1):pp. 87–94. 53

Fleischmann, R. D., Alland, D., Eisen, J. A., Carpenter, L., White, O., Peterson, J., DeBoy, R., Dodson, R., Gwinn, M., Haft, D., Hickey, E., Kolonay, J. F., Nelson, W. C., Umayam, L. A., Ermolaeva, M., Salzberg, S. L., Delcher, A., Utterback, T., Weidman, J., Khouri, H., Gill, J., Mikula, A., Bishai, W., Jr, W. R. J., Venter, J. C., and Fraser, C. M. (2002). Whole-genome comparison of mycobacterium tuberculosis clinical and laboratory strains. *J Bacteriol*, 184(19):5479–5490. 36, 90

Fraser, H. B., Hirsh, A. E., Wall, D. P., and Eisen, M. B. (2004). Coevolution of gene expression among interacting proteins. *Proc Natl Acad Sci U S A*, 101(24):9033–9038. 19, 24, 50

Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social Networks*, 1:215–239. 46

Fryxell, K. J. (1996). The coevolution of gene family trees. *Trends Genet*, 12(9):364–369. 20, 21

Ge, H., Liu, Z., Church, G. M., and Vidal, M. (2001). Correlation between transcriptome and interactome mapping data from saccharomyces cerevisiae. *Nat Genet*, 29(4):482–486. 19

Göbel, U., Sander, C., Schneider, R., and Valencia, A. (1994). Correlated mutations and residue contacts in proteins. *Proteins*, 18(4):309–317. 21

Goh, C. S., Bogan, A. A., Joachimiak, M., Walther, D., and Cohen, F. E. (2000). Coevolution of proteins with their interaction partners. *J Mol Biol*, 299(2):283–293. 26

Gomez, S. M., Noble, W. S., and Rzhetsky, A. (2003). Learning to predict protein-protein interactions from protein sequences. *Bioinformatics*, 19(15):1875–1881. 29

Hancock, R. E. and Nikaido, H. (1978). Outer membranes of gram-negative bacteria. xix. isolation from pseudomonas aeruginosa pao1 and use in reconstitution and definition of the permeability barrier. *J Bacteriol*, 136(1):381–390. 9

He, X. and Zhang, J. (2006). Why do hubs tend to be essential in protein networks? *PLoS Genet*, 2(6):e88. 46, 63

Ho, Y., Gruhler, A., Heilbut, A., Bader, G. D., Moore, L., Adams, S.-L., 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, A. R., Sassi, H., Nielsen, P. A., Rasmussen, K. J., Andersen, J. R., Johansen, L. E., Hansen, L. H., Jespersen, H., Podtelejnikov, A., Nielsen, E., Crawford, J., Poulsen, V., SÃ¸rensen, B. D., Matthiesen, J., Hendrickson, R. C., Gleeson, F., Pawson, T., Moran, M. F., Durocher, D., Mann, M., Hogue, C. W. V., Figeys, D., and Tyers, M. (2002). Systematic identification of protein complexes in saccharomyces cerevisiae by mass spectrometry. *Nature*, 415(6868):180–183. 17

Hunter, S., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Das, U., Daugherty, L., Duquenne, L., Finn, R. D., Gough, J., Haft, D., Hulo, N., Kahn, D., Kelly, E., Laugraud, A., Letunic, I., Lonsdale, D., Lopez, R., Madera, M., Maslen, J., McAnulla, C., McDowall, J., Mistry, J., Mitchell, A., Mulder, N., Natale, D., Orengo, C., Quinn, A. F., Selengut, J. D., Sigrist, C. J. A., Thimma, M., Thomas, P. D., Valentin, F., Wilson, D., Wu, C. H., and Yeats, C. (2009). Interpro: the integrative protein signature database. *Nucleic Acids Res*, 37(Database issue):D211–D215. 33

Huntley, R. P., Binns, D., Dimmer, E., Barrell, D., O'Donovan, C., and Apweiler, R. (2009). Quickgo: a user tutorial for the web-based gene ontology browser. *Database (Oxford)*, 2009:bap010. 42

Ivchenko, G. and Honov, S. A. (1998). On the jaccard similarity test. *Journal of Mathematical Sciences*, 88. Jaccard index calculation. 49

Jain, S. and Bader, G. D. (2010). An improved method for scoring protein-protein interactions using semantic similarity within the gene ontology. *BMC Bioinformatics*, 11:562. 50

Jasmer, R. M., Nahid, P., and Hopewell, P. C. (2002). Clinical practice. latent tuberculosis infection. *N Engl J Med*, 347(23):1860–1866. 1

Kenri, T., Seto, S., Horino, A., Sasaki, Y., Sasaki, T., and Miyata, M. (2004). Use of fluorescent-protein tagging to determine the subcellular localization of mycoplasma pneumoniae proteins encoded by the cytadherence regulatory locus. *J Bacteriol*, 186(20):6944–6955. 9

Keskin, O., Gursoy, A., Ma, B., and Nussinov, R. (2008). Principles of protein-protein interactions: what are the preferred ways for proteins to interact? *Chem Rev*, 108(4):1225–1244. 11, 12, 14

King, M. W. (2011). Signal transduction. http://themedicalbiochemistrypage.org/signal-transduction.html. 92

Koch, R. (1905). Nobel prize winning lecture. http://nobelprize.org/nobel_prizes/medicine/laureates/1905/koch-lecture.html. Award winning lecture. 2

Kotelnikova, E., Kalinin, A., Yuryev, A., and Maslov, S. (2007). Prediction of protein-protein interactions on the basis of evolutionary conservation of protein functions. *Evol Bioinform Online*, 3:197–206. 28

Kumar, R. B., Xie, Y. H., and Das, A. (2000). Subcellular localization of the agrobacterium tumefaciens t-dna transport pore proteins: Virb8 is essential for the assembly of the transport pore. *Mol Microbiol*, 36(3):608–617. 9

Kundrotas, P. J. and Alexov, E. (2007). Protcom: searchable database of protein complexes enhanced with domain-domain structures. *Nucleic Acids Res*, 35(Database issue):D575–D579. 31

Marcotte, E. M., Pellegrini, M., Ng, H. L., Rice, D. W., Yeates, T. O., and Eisenberg, D. (1999). Detecting protein function and protein-protein interactions from genome sequences. *Science*, 285(5428):751–753. 17, 18, 28, 29

Mazandu, G. K. and Mulder, N. J. (2011). Go-universal metric for measuring term closeness in gene ontology (go). 51

Meur, N. L. and Gentleman, R. (2008). Modeling synthetic lethality. *Genome Biol*, 9(9):R135. 17

Mika, S. and Rost, B. (2006). Protein-protein interactions more conserved within species than across species. *PLoS Comput Biol*, 2(7):e79. 11

Mistry, M. and Pavlidis, P. (2008). Gene ontology term overlap as a measure of gene functional similarity. *BMC Bioinformatics*, 9:327. 49

Moya, A., Peretó, J., Gil, R., and Latorre, A. (2008). Learning how to live together: genomic insights into prokaryote-animal symbioses. *Nat Rev Genet*, 9(3):218–229. 24

Mueller, T. D. and Feigon, J. (2002). Solution structures of uba domains reveal a conserved hydrophobic surface for protein-protein interactions. *J Mol Biol*, 319(5):1243–1255. 13

Ng, S.-K., Zhang, Z., and Tan, S.-H. (2003a). Integrative approach for computationally inferring protein domain interactions. *Bioinformatics*, 19(8):923–929. 30, 31

Ng, S.-K., Zhang, Z., Tan, S.-H., and Lin, K. (2003b). Interdom: a database of putative interacting protein domains for validating predicted protein interactions and complexes. *Nucleic Acids Res*, 31(1):251–254. 8, 13, 29, 30, 87

Oliver, S. (2000). Guilt-by-association goes global. *Nature*, 403(6770):601–603. 12, 16

Pagel, P., Oesterheld, M., Tovstukhina, O., Strack, N., Stümpflen, V., and Frishman, D. (2008). Dima 2.0–predicted and known domain interactions. *Nucleic Acids Res*, 36(Database issue):D651–D655. 30, 88

Pagès, S., Bélaïch, A., Bélaïch, J. P., Morag, E., Lamed, R., Shoham, Y., and Bayer, E. A. (1997). Species-specificity of the cohesin-dockerin interaction between clostridium thermocellum and clostridium cellulolyticum: prediction of specificity determinants of the dockerin domain. *Proteins*, 29(4):517–527. 20

Pang, K., Sheng, H., and Ma, X. (2010). Understanding gene essentiality by finely characterizing hubs in the yeast protein interaction network. *Biochem Biophys Res Commun*, 401(1):112–116. 46, 63

Pazos, F., Helmer-Citterich, M., Ausiello, G., and Valencia, A. (1997). Correlated mutations contain information about protein-protein interaction. *J Mol Biol*, 271(4):511–523. 21

Pazos, F., Juan, D., Izarzugaza, J. M. G., Leon, E., and Valencia, A. (2008). Prediction of protein interaction based on similarity of phylogenetic trees. *Methods Mol Biol*, 484:523–535. 18

Pazos, F. and Valencia, A. (2001). Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Eng*, 14(9):609–614. 21

Pazos, F. and Valencia, A. (2008). Protein co-evolution, co-adaptation and interactions. *EMBO J*, 27(20):2648–2655. 24, 26

Pieper, U., Webb, B. M., Barkan, D. T., Schneidman-Duhovny, D., Schlessinger, A., Braberg, H., Yang, Z., Meng, E. C., Pettersen, E. F., Huang, C. C., Datta, R. S., Sampathkumar, P., Madhusudhan, M. S., SjÃ¶lander, K., Ferrin, T. E., Burley, S. K., and Sali, A. (2011). Modbase, a database of annotated comparative protein structure models, and associated resources. *Nucleic Acids Res*, 39(Database issue):D465–D474. 13

Plotkin, J. B., Dushoff, J., and Fraser, H. B. (2004). Detecting selection using a single genome sequence of m. tuberculosis and p. falciparum. *Nature*, 428(6986):942–945. 25, 48

Pond, S. L. K., Frost, S. D. W., and Muse, S. V. (2005). Hyphy: hypothesis testing using phylogenies. *Bioinformatics*, 21(5):676–679. 25, 47

Prahlad T. Ram, Ravi Iyengar (2008). Signal transduction. http://www.accessscience.com. 92

Puig, O., Caspary, F., Rigaut, G., Rutz, B., Bouveret, E., Bragado-Nilsson, E., Wilm, M., and Séraphin, B. (2001). The tandem affinity purification (tap) method: a general procedure of protein complex purification. *Methods*, 24(3):218–229. 2

Raghavachari, B., Tasneem, A., Przytycka, T. M., and Jothi, R. (2008). Domine: a database of protein domain interactions. *Nucleic Acids Res*, 36(Database issue):D656–D661. 29, 30, 88

Ramos-Vara, J. A. (2005). Technical aspects of immunohistochemistry. *Vet Pathol*, 42(4):405–426. 10

Resnik (1995). Using information content to evaluate semantic similarity in a taxonomy. In *14th International Joint Conference on Artificial Intelligence*, pages 448–453. 50

Rey, S., Gardy, J. L., and Brinkman, F. S. L. (2005). Assessing the precision of high-throughput computational and laboratory approaches for the genome-wide identification of protein subcellular localization in bacteria. *BMC Genomics*, 6:162. 9, 10

Richardson, J. S. (1981). The anatomy and taxonomy of protein structure. *Adv Protein Chem*, 34:167–339. 8

Rigaut, G., Shevchenko, A., Rutz, B., Wilm, M., Mann, M., and Sèraphin, B. (1999). A generic protein purification method for protein complex characterization and proteome exploration. *Nat Biotechnol*, 17(10):1030–1032. 13, 17

Rincón-Limas, D. E., Lu, C. H., Canal, I., Calleja, M., Rodríguez-Esteban, C., Izpisùa-Belmonte, J. C., and Botas, J. (1999). Conservation of the expression and function of apterous orthologs in drosophila and mammals. *Proc Natl Acad Sci U S A*, 96(5):2165–2170. 35

Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*, 4(4):406–425. 48

Sanz, J., Navarro, J., ArbuÃ©s, A., MartÃn, C., MarijuÃ¡n, P. C., and Moreno, Y. (2011). The transcriptional regulatory network of mycobacterium tuberculosis. *PLoS One*, 6(7):e22178. 83

Schlicker, A. and Albrecht, M. (2008). Funsimmat: a comprehensive functional similarity database. *Nucleic Acids Res*, 36(Database issue):D434–D439. 42

Schuster, S. C. (2008). Next-generation sequencing transforms today's biology. *Nat Methods*, 5(1):16–18. 3

Scott, M. S., Calafell, S. J., Thomas, D. Y., and Hallett, M. T. (2005). Refining protein subcellular localization. *PLoS Comput Biol*, 1(6):e66. 9

Sevilla, J. L., Segura, V., Podhorski, A., Guruceaga, E., Mato, J. M., Martínez-Cruz, L. A., Corrales, F. J., and Rubio, A. (2005). Correlation between gene expression and go semantic similarity. *IEEE/ACM Trans Comput Biol Bioinform*, 2(4):330–338. 50

Sharan, R., Suthram, S., Kelley, R. M., Kuhn, T., McCuine, S., Uetz, P., Sittler, T., Karp, R. M., and Ideker, T. (2005). Conserved patterns of protein interaction in multiple species. *Proc Natl Acad Sci U S A*, 102(6):1974–1979. 88

Shevchenko, A., Jensen, O. N., Podtelejnikov, A. V., Sagliocco, F., Wilm, M., Vorm, O., Mortensen, P., Shevchenko, A., Boucherie, H., and Mann, M. (1996). Linking genome and proteome by mass spectrometry: large-scale identification of yeast proteins from two dimensional gels. *Proc Natl Acad Sci U S A*, 93(25):14440–14445. 10

Shoemaker, B. A. and Panchenko, A. R. (2007a). Deciphering protein-protein interactions. part i. experimental techniques and databases. *PLoS Comput Biol*, 3(3):e42. 2

Shoemaker, B. A. and Panchenko, A. R. (2007b). Deciphering ProteinâProtein Interactions. Part I. Experimental Techniques and Databases. *PLoS Computational Biology*, 3(3). 11, 12, 13, 15

Sikder, A. R. and Zomaya, A. Y. (2008). Inferring boundary information of discontinuous-domain proteins. *IEEE Trans Nanobioscience*, 7(3):200–205. 8, 9

Skrabanek, L., Saini, H. K., Bader, G. D., and Enright, A. J. (2008). Computational prediction of protein-protein interactions. *Mol Biotechnol*, 38(1):1–17. 11, 12, 23

Sprinzak, E. and Margalit, H. (2001). Correlated sequence-signatures as markers of protein-protein interaction. *J Mol Biol*, 311(4):681–692. 18, 28

Stein, A., Russell, R. B., and Aloy, P. (2005). 3did: interacting protein domains of known three-dimensional structure. *Nucleic Acids Res*, 33(Database issue):D413–D417. 30

Sussman, J. L., Lin, D., Jiang, J., Manning, N. O., Prilusky, J., Ritter, O., and Abola, E. E. (1998). Protein data bank (pdb): database of three-dimensional structural information of biological macromolecules. *Acta Crystallogr D Biol Crystallogr*, 54(Pt 6 Pt 1):1078–1084. 88

Teichmann, S. A. (2002). The constraints protein-protein interactions place on sequence divergence. *J Mol Biol*, 324(3):399–407. 90, 94

Thompson, J. N. (1989). Concepts of coevolution. *Trends Ecol Evol*, 4(6):179–183. 24

Tsoka, S. and Ouzounis, C. A. (2000). Prediction of protein interactions: metabolic enzymes are frequently involved in gene fusion. *Nat Genet*, 26(2):141–142. 18

Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., 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., and Rothberg, J. M. (2000). A comprehensive analysis of protein-protein interactions in saccharomyces cerevisiae. *Nature*, 403(6770):623–627. 13

Valdar, W. S. and Thornton, J. M. (2001). Protein-protein interfaces: analysis of amino acid conservation in homodimers. *Proteins*, 42(1):108–124. 11, 13, 90

Valencia, A. and Pazos, F. (2002). Computational methods for the prediction of protein interactions. *Curr Opin Struct Biol*, 12(3):368–373. 11, 17, 18, 19, 21, 22, 23

von Mering, C., Huynen, M., Jaeggi, D., Schmidt, S., Bork, P., and Snel, B. (2003a). String: a database of predicted functional associations between proteins. *Nucleic Acids Res*, 31(1):258–261. 28, 55, 56

von Mering, C., Huynen, M., Jaeggi, D., Schmidt, S., Bork, P., and Snel, B. (2003b). String: a database of predicted functional associations between proteins. *Nucleic Acids Res*, 31(1):258–261. 39

von Mering, C., Jensen, L. J., Kuhn, M., Chaffron, S., Doerks, T., Krüger, B., Snel, B., and Bork, P. (2007). String 7–recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res*, 35(Database issue):D358–D362. 3, 12, 14, 28, 39, 46, 56, 89

von Mering, C., Jensen, L. J., Snel, B., Hooper, S. D., Krupp, M., Foglierini, M., Jouffre, N., Huynen, M. A., and Bork, P. (2005). String: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res*, 33(Database issue):D433–D437. 40, 45, 46

von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S., and Bork, P. (2002). Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 417(6887):399–403. 14

Walhout, A. J., Sordella, R., Lu, X., Hartley, J. L., Temple, G. F., Brasch, M. A., Thierry-Mieg, N., and Vidal, M. (2000). Protein interaction mapping in c. elegans using proteins involved in vulval development. *Science*, 287(5450):116–122. 35, 87

Wang, J., Zhou, X., Zhu, J., Zhou, C., and Guo, Z. (2010). Revealing and avoiding bias in semantic similarity scores for protein pairs. *BMC Bioinformatics*, 11:290. 49

Wang, J. Z., Du, Z., Payattakool, R., Yu, P. S., and Chen, C.-F. (2007). A new method to measure the semantic similarity of go terms. *Bioinformatics*, 23(10):1274–1281. 42

Weiner, J. and Bornberg-Bauer, E. (2006). Evolution of circular permutations in multidomain proteins. *Mol Biol Evol*, 23(4):734–743. 28

Wetlaufer, D. B. (1973). Nucleation, rapid folding, and globular intrachain regions in proteins. *Proc Natl Acad Sci U S A*, 70(3):697–701. 8

WHO (2009). *Tuberculosis fact sheet*. 1, 2

Wojcik, J. and Schächter, V. (2001). Protein-protein interaction map inference using interacting domain profile pairs. *Bioinformatics*, 17 Suppl 1:S296–S305. 29

Xenarios, I., Rice, D. W., Salwinski, L., Baron, M. K., Marcotte, E. M., and Eisenberg, D. (2000). Dip: the database of interacting proteins. *Nucleic Acids Res*, 28(1):289–291. 14

Yang, Z. (1998). Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol Biol Evol*, 15(5):568–573. 25

Yeang, C.-H. (2008). Identifying coevolving partners from paralogous gene families. *Evol Bioinform Online*, 4:97–107. 24, 25

Yellaboina, S., Dudekula, D. B., and Ko, M. S. (2008). Prediction of evolutionarily conserved interologs in mus musculus. *BMC Genomics*, 9:465. 87, 88

Young, K. H. (1998). Yeast two-hybrid: so many interactions, (in) so little time... *Biol Reprod*, 58(2):302–311. 11, 13

Young, L., Jernigan, R. L., and Covell, D. G. (1994). A role for surface hydrophobicity in protein-protein recognition. *Protein Sci*, 3(5):717–729. 13

Yu, H., Luscombe, N. M., Lu, H. X., Zhu, X., Xia, Y., Han, J.-D. J., Bertin, N., Chung, S., Vidal, M., and Gerstein, M. (2004). Annotation transfer between genomes: protein-protein interologs and protein-dna regulogs. *Genome Res*, 14(6):1107–1118. 36, 88

Yu, H., Paccanaro, A., Trifonov, V., and Gerstein, M. (2006). Predicting interactions in protein networks by completing defective cliques. *Bioinformatics*, 22(7):823–829. 23

Yu, N. Y., Wagner, J. R., Laird, M. R., Melli, G., Rey, S., Lo, R., Dao, P., Sahinalp, S. C., Ester, M., Foster, L. J., and Brinkman, F. S. L. (2010). Psortb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics*, 26(13):1608–1615. 10

Zhu, H., Klemic, J. F., Chang, S., Bertone, P., Casamayor, A., Klemic, K. G., Smith, D., Gerstein, M., Reed, M. A., and Snyder, M. (2000). Analysis of yeast protein kinases using protein chips. *Nat Genet*, 26(3):283–289. 11, 13

Zotenko, E., Mestre, J., O'Leary, D. P., and Przytycka, T. M. (2008). Why do hubs in the yeast protein interaction network tend to be essential: reexamining the connection between the network topology and essentiality. *PLoS Comput Biol*, 4(8):e1000140. 46, 63