Computational Studies of Host-Pathogen Protein-Protein Interactions—A case study of the *H. sapiens — M. tuberculosis* H37Rv system

Hufeng Zhou

(*B.A, HUST*)

(*B.E, HZAU*)

A Thesis submitted for the degree of

Doctor of Philosophy

NUS Graduate School for Integrative Sciences and Engineering

National University of Singapore

2013

Declaration

I hereby declare that this thesis is my original work and it has been written by me in its entirety.
I have duly acknowledged all the source of information which have been used in this thesis.

Hufeng Zhou

30 April 2013

Acknowledgements

First and foremost, I would like to express my immense gratitude to my supervisor Professor Limsoon Wong. He helped me successfully make the transition from being an experimental biologist to become a competent computational biologist and initiated my academic journey. Over the past few years, I have benefited tremendously from his excellent guidance, persistent support, and invaluable advice. Working with him was extremely pleasant. I have learnt a lot from him in many aspects of doing research. His enthusiasm, dedication and preciseness have deeply influenced me.

I want to thank my family. I am deeply indebted to my parents Hongcao Zhou and Lifang Hu for their unconditional love, understanding and support. Their love and support are the source of motivation and happiness in my life.

Finally, I appreciate the friendship and support of our current and former group members: Jingjing Jin, Chern-Han Yong, Dr. Liu Bing, Dr. Difeng Dong, Dr. Tsung-Han Chiang, Mengyuan Fan, Michal Wozniak, Junliang Kevin Lim and many others. I would like to express my sincerest gratitude to them for the collaborative and friendly environment as well as the countless useful discussions.

Contents

Summary

Host–pathogen protein-protein interaction (PPI) data are very important information for illuminating infection mechanisms and for developing better prevention measures.

However, host–pathogen PPI data are very scarce in most host–pathogen systems. Computational prediction of host–pathogen PPIs is an important strategy to fill in the gap. In this dissertation, we systemically investigate host–pathogen protein-protein interactions using the *H. sapiens–M. tuberculosis* H37Rv system as the model host–pathogen system. Our four main contributions are summarized below.

Knowledge of intra-species PPIs could help a lot in understanding the functional role of the proteins that are involved in host–pathogen PPIs. Moreover, intra-species pathogen PPIs have been used as training data for the prediction of host–pathogen PPIs(Dyer et al., 2007). But for most pathogens, their intra-species pathogen PPIs are not readily available on a large scale; this is especially true for *M. tuberculosis* H37Rv. Therefore, in Chapter 2, we identify a reliable *M. tuberculosis* H37Rv PPI dataset and pave the way for the analysis of *H. sapiens–M. tuberculosis* H37Rv PPIs.

For most host–pathogen systems, including *H. sapiens–M. tuberculosis* H37Rv, high-quality large-scale inter-species PPIs are scarce, resulting in a lack of gold standard to assess the predicted host–pathogen PPIs. Therefore, functional analysis based on pathway data becomes one of the most frequently used approaches to assess the predicted host–pathogen PPIs. However, there are several major limitations that seriously reduce the effective use of pathway data for analysis and assessment of predicted host–pathogen PPIs. Thus, in Chapter 3 we create an analysis tool, IntPath, which is currently one of the most comprehensive pathway integration databases. IntPath enables comprehensive functional analysis based on integrated pathway data for both host and pathogen. It uses a novel integration technology that addresses limitations of current pathway databases; and it also provides the scalability to extend to many model host organisms and important pathogens.

Domain-domain interaction (DDI) based approaches are often used for predicting

both intra-species and inter-species PPIs, with the assumption that domain-domain interactions mediate the protein-protein interactions. In Chapter 4, we develop an accurate DDI-based prediction approach with emphasis on (i) differences between the specific domain sequences on annotated regions of proteins under the same domain ID and (ii) calculation of the interaction strength of predicted PPIs based on the interacting residues in their interaction interfaces. We compare our accurate DDI-based approach to a conventional DDI-based approach for predicting PPIs based on gold standard intra-species PPIs and coherent informative Gene Ontology assessment. The assessment results show that our accurate DDI-based approach achieves much better performance in predicting PPIs than the convention approach.

Homology-based approaches are also used in predicting host–pathogen PPIs in many works, but with unsolved deficiencies in the transfer of interactions from template PPIs. In Chapter 5, we develop an accurate homology-based prediction approach by taking into account (i) differences between eukaryotic and prokaryotic proteins and (ii) differences between inter-species and intra-species PPI interfaces. We compare our accurate homology-based approach to a conventional homology-based approach for predicting host–pathogen PPIs based on cellular compartment distribution analysis, disease gene list enrichment analysis, pathway enrichment analysis and functional category enrichment analysis. The analysis results support the validity of our prediction result and clearly show that our accurate homology-based approach has better performance in predicting *H. sapiens–M. tuberculosis* H37Rv PPIs.

List of Figures

List of Tables

Chapter 1

Introduction and Background

Host-pathogen interactions are important for understanding infection mechanism and developing better treatment and prevention of infectious diseases. The protein interaction map will guide the investigation on the key PPIs that may lead to the adhesion, colonization, and even invasion of pathogens to human cells. However, prediction of host-pathogen PPIs has its unique challenges.

Many approaches for predicting intra-species PPIs may not be applicable to inter-species host-pathogen PPIs. For example, if two interacting partners are located at the same cellular compartment, they are more likely to interact with each other in the intra-species scenario, because being at the same cellular compartment (i.e., being in the same place) is a requirement for interaction. But this is inapplicable to host-pathogen PPIs: The cellular compartment annotations for host (resp. pathogen) proteins refer to cellular compartments in the host (resp. pathogen) species and, thus, the host and pathogen proteins in a host-pathogen PPI are never annotated for the same cellular compartment. Therefore novel computational prediction and assessment approaches are needed for the study of inter-species host-pathogen PPIs.

Many computational studies on host-pathogen interactions have been published. Here, we first review recent progress and results in this field, providing a system-

atic summary, comparison and discussion of computational studies on host-pathogen interactions including: prediction and analysis of host-pathogen protein-protein interactions; basic principles revealed from host-pathogen interactions; and database and software tools for host-pathogen interaction data collection, integration and analysis. After the review, we state the objectives of this dissertation and highlight our main results.

1.1 Context and introduction

Infectious diseases are among the leading causes of death especially in the developing world. Host-pathogen interactions are crucial for better understanding of the mechanisms that underlie infectious diseases and for developing more effective treatment and prevention measures.

While host-pathogen interactions take many forms, in this review, we concentrate on protein-protein interactions (PPIs) between a pathogen and its host. This Chapter consists of the following parts: (i) host-pathogen PPI prediction; (ii) basic principles derived from analysis of known host-pathogen PPIs; (iii) host-pathogen PPI analysis and assessment; and (iv) host-pathogen interaction data collection and integration.

Several approaches have been proposed to computationally predict host-pathogen protein-protein interactions. There has also been progress on analyzing and assessing the quality of the inferred host-pathogen PPIs. This has led to cataloging of PPI data that can be further analyzed to understand the impact of these interactions (especially on the host) and to decipher the underlying disease mechanisms. Approaches developed for predicting host-pathogen PPIs can be broadly categorized into homology-based(Lee et al., 2008; Krishnadev and Srinivasan, 2008; Tyagi et al., 2009; Krishnadev and Srinivasan, 2011; Wuchty, 2011), structure-based(Davis et al., 2007; Doolittle and Gomez, 2011, 2010), domain and motif interaction-based approaches(Dyer et al., 2007; Evans et al., 2009), as well as machine learning-based approaches(Tastan et al., 2009; Dyer

et al., 2011; Qi et al., 2010). These approaches can also be combined and used together in some studies to improve prediction performance. These approaches are reviewed in Section1.2 "Host-pathogen protein-protein interactions prediction".

An analysis of experimentally verified, as well as manually curated, host-pathogen PPIs have led to a number of observations. These observations include the topological properties of targeted host proteins and structural properties of host-pathogen protein-protein interaction interfaces. These observations are discussed in Section 1.3 "Basic principles of host-pathogen interaction".

Approaches for assessing and analyzing host-pathogen PPIs can be categorized into assessment based on gold standard PPIs(Tastan et al., 2009; Qi et al., 2010; Dyer et al., 2011; Evans et al., 2009; Davis et al., 2007; Doolittle and Gomez, 2011); functional information analysis in terms of Gene Ontology(Davis et al., 2007; Wuchty, 2011; Tastan et al., 2009; Doolittle and Gomez, 2010, 2011; Evans et al., 2009), pathways(Singh et al., 2010; Zhao et al., 2011; Wuchty, 2011; Evans et al., 2009), gene expression data(Wuchty, 2011; Krishnadev and Srinivasan, 2008; Davis et al., 2007) and RNA interference data(Doolittle and Gomez, 2010, 2011; Evans et al., 2009; Tastan et al., 2009; Qi et al., 2010; Dyer et al., 2011); localization information analysis in terms of protein sub-cellular localization(Lee et al., 2008; Krishnadev and Srinivasan, 2008; Tyagi et al., 2009; Krishnadev and Srinivasan, 2011; Wuchty, 2011) and co-localization of host and pathogen proteins(Doolittle and Gomez, 2011, 2010); related experimental data analyses(Doolittle and Gomez, 2010; Tastan et al., 2009; Qi et al., 2010); and biological case studies and explanations(Krishnadev and Srinivasan, 2008; Tyagi et al., 2009; Krishnadev and Srinivasan, 2011; Dyer et al., 2011; Davis et al., 2007; Doolittle and Gomez, 2011, 2010). Some of these assessment approaches can also be used as filtering strategies for pruning host-pathogen PPI prediction results. These approaches and the outcome of the analysis are reviewed in Section 1.4 "Analysis and assessment of host-pathogen PPIs".

Host-pathogen PPIs curated from primary literature are usually facilitated by text-mining techniques(Chatr-aryamontri et al., 2009; Navratil et al., 2009). With more host-pathogen PPI data available from literature curation and experiments, there are strong needs for data collection and integration facilities that can provide comprehensive storage, convenient access, and effective analysis of the integrated host-pathogen interaction data. The development of software and database tools dedicated to host-pathogen interaction data collection, integration and analysis are also very prominent. Integration of host-pathogen interaction data are not confined to PPI data. Other related data — such as pathogen virulence factors, human diseases related genes, sequence and homology information, pathway information, functional annotations, diseases information, and literature sources, etc.—are also being integrated into several databases. These databases (Winnenburg et al., 2008; Fu et al., 2009; Chatr-aryamontri et al., 2009; Navratil et al., 2009; Xiang et al., 2007; Ranjit and Bindu, 2010; Fahey et al., 2011; Driscoll et al., 2009, 2011; Gillespie et al., 2011) and softwares(Sergey et al., 2011) are reviewed in Section 1.5 “Host-pathogen interaction data collection and integration”.

1.2 Host-pathogen protein-protein interactions prediction

Host-pathogen protein-protein interactions play an important role between the host and pathogen, which may be crucial in the outcome of an infection and the establishment of disease. Unfortunately, experimentally verified interactions between host and pathogen proteins are currently rather limited for most host-pathogen systems. This has motivated a number of pioneering works on computational prediction of host-pathogen protein-protein interactions. These works can be roughly categorized into modeling approaches based on sequence homology, protein structure, domain and motif, and approaches based on machine learning. These pioneering works are reviewed and discussed below.

1.2.1 Homology-based approach

The homology-based approach is a conventional way for predicting intra-species PPIs. Many studies have also adopted this strategy for predicting host-pathogen PPIs, which are inter-species PPIs. The basic hypothesis of the homology-based approach is that the interaction between a pair of proteins in one species is expected to be conserved in related species(Matthews et al., 2001). This is a reasonable hypothesis as a pair of homologous proteins are descended from the same ancestral pair of interacting proteins and is expected to inherit the structure and function and, thus, interactions of the ancestral proteins. Therefore, the basic procedure of the homology-based approach for intra-species PPI prediction is to (i) start from a known PPI (the template PPI) in some source species, (ii) determining in the target species the homologs (x', y') of the two proteins (x, y) in the template PPI, and (iii) predicting that the two homologs (x', y') interact in the target species. This approach is generally adapted to the inter-species scenario of host-pathogen PPI prediction by (i) starting from a known PPI (the template PPI) in some source species, (ii) determining in the host a homolog (x') and in the pathogen a homolog (y') respectively of the two proteins (x,y) in the template PPI, and (iii) predicting that (x',y') interact.

The main advantages of the homology-based approach to host-pathogen PPI prediction are its simplicity and its apparent biological basis. Since the data required for performing the prediction are only the template PPIs and protein sequences, this approach is scalable and can be applied to many different host-pathogen systems. The homology-based approach can be used alone(Lee et al., 2008; Krishnadev and Srinivasan, 2008; Tyagi et al., 2009; Krishnadev and Srinivasan, 2011) or in combination with other methods(Wuchty, 2011) in predicting host-pathogen PPIs. The investigated host-pathogen systems in past studies include *H. sapiens–P. falciparum*(Wuchty, 2011; Lee et al., 2008; Krishnadev and Srinivasan, 2008), *H. sapiens–H. pylori*(Tyagi et al., 2009), *E. coli–phage T4*(Krishnadev and Srinivasan, 2011), *E. coli–phage lambda*(Krishnadev and

Srinivasan, 2011), *H. sapiens–E. coli*(Krishnadev and Srinivasan, 2011), *H. sapiens–S. enterica*(Krishnadev and Srinivasan, 2011), *H. sapiens–Y. pestis*(Krishnadev and Srinivasan, 2011), etc. The template PPIs used in the prediction can also be very different. The commonly used template PPIs are from DIP(Salwinski et al., 2004), iPfam(Finn et al., 2005), MINT(Zanzoni et al., 2002), HPRD(Mishra et al., 2006), Reactome(Joshi-Tope et al., 2005), IntAct(Hermjakob et al., 2004), etc.

There is an inherent weakness in the homology-based approach. Basically, in a real biological process, such as infection, the two proteins in a predicted PPI may actually have little opportunity to be present together. Consequently, host-pathogen PPIs predicted solely on the homology basis, without considering other biological properties of the proteins involved, may not be very reliable. Additional information should be used to increase the accuracy of the prediction. For example, extracellular localization and trans-membrane regions are used in pruning(Krishnadev and Srinivasan, 2011) or constraining the predictions(Tyagi et al., 2009). Also, a pathogen (e.g., *P. falciparum*) may infect different organs at different stages of the pathogen's life cycle. Thus, filtering by tissue-specific gene expression data may also improve prediction reliability(Krishnadev and Srinivasan, 2008). Indeed, recognizing this weakness in the homology-based approach, Wuchty (2011) has proposed filtering PPIs predicted by the homology-based approach using a random-forest classifier trained on sequence compositional characteristics of known PPIs, as well as by gene expression and molecular characteristics. This results in a significantly smaller set of putative host-pathogen PPIs, which are claimed to be of higher quality than the original set of predicted PPIs.

1.2.2 Structure-based approach

When a pair of proteins have structures that are similar to a known interacting pair of proteins, it is reasonable to believe that the former are likely interacting in a way that is structurally similar to the latter. In accordance to this hypothesis, several works

have used structural information to identify the similarity between query proteins (i.e., proteins in the pathogen and host) and template PPIs (i.e., known interacting protein pairs), and infer that host-pathogen protein pairs that match some template PPIs are interacting.

Comparative modeling

Prediction by comparative modeling is a representative structure-based approach. For example, in Davis et al. (2007), an automated pipeline for large-scale comparative protein structure modeling, MODPIPE, is applied to model the structure of host and pathogen proteins based on their sequences and corresponding template structures. Given the computed model of a protein, the SCOP(Murzin et al., 1995) superfamilies that the protein belongs to are identified. A database of protein structural interfaces, PIBASE, is then scanned. If a SCOP superfamily of a host protein and a SCOP superfamily of a pathogen protein are both involved in the same PIBASE(Davis and Sali, 2005) protein structural interface, then the host protein and the pathogen protein are predicted as a putative PPI.

Query proteins that lack structural templates cannot be modeled in the process above. In this case, template interactions in alternative databases (e.g., IntAct) are considered by Davis et al. (2007). Specifically, a pair of host and pathogen proteins are predicted to interact if at least 50% of each of the two protein sequences are similar to some member proteins of a template complex in IntAct and the joint sequence identity ($\sqrt{Sequence\ Identity1 * Sequence\ Identity2}$) is at least 80%. These predictions, which are conducted without structural information, form a very small portion of the total number of putative PPIs, because of the stringent joint threshold. Each prediction is further followed by a series of assessments and filtering (biological and network filters), which results in a significant reduction of potential host-pathogen PPIs by several order of magnitudes.

Structural similarity

Structural similarity can also be analyzed using the Dali database(Holm et al., 2008). This strategy has been adopted to predict *H. sapiens*–HIV PPIs(Doolittle and Gomez, 2010), *H.sapiens*–DENV PPIs(Doolittle and Gomez, 2011), and *A. aegypti*–DENV PPIs(Doolittle and Gomez, 2011). Dali calculates structural similarity score by comparing the 3D structural coordinates of two PDB entries(Doolittle and Gomez, 2011). To predict the *H. sapiens*–HIV and *H. sapiens*–DENV PPIs, structurally similar pathogen (HIV, DENV) and host (*H. sapiens*) proteins are first determined using Dali. Then, under the assumption that pathogen proteins having similar structure to host proteins are likely to participate in a similar set of PPIs (*H. sapiens* PPI dataset from HPRD(Mishra et al., 2006)) that those matched host proteins participate in, the pathogen proteins are directly mapped to their high-similarity matches within the host intra-species PPI network to predict the host-pathogen PPIs(Doolittle and Gomez, 2010, 2011). The same structural similarity prediction method has been applied to identify orthologs between *D. melanogaster* and *A. aegypti* and map *D. melanogaster*–DENV PPIs to predict *A. aegypti*–DENV PPIs(Doolittle and Gomez, 2011)—the host-pathogen PPIs between DENV and its real insect host. The accuracy of this prediction method depends on the performance of Dali in determining structurally similar pathogen and host proteins. The availability of pathogen and host protein structures and the quality of host intra-species PPI data also have a significant influence on prediction results.

1.2.3 Domain and motif interaction-based approach

Domains are basic building blocks determining the structure and function of proteins and they play specialized role in mediating the interaction of proteins with other molecules(Itzhaki et al., 2010). Some studies have proposed predicting host-pathogen PPI based on domain-domain interaction (DDI)(Dyer et al., 2007) and motif-domain interaction(Evans et al., 2009).

Domain-domain interaction-based approach

Dyer et al. (2007) predict host-pathogen PPIs in the *H. sapiens–P. falciparum* system by integrating known intra-species PPIs with domain profiles based on an association method (sequence-signature algorithm) proposed by Sprinzak and Margalit (2001). Specifically, domains are first identified by InterProScan(Quevillon et al., 2005) in each interacting protein in the intra-species PPIs. Then, the probability $P(d, e)$ that two proteins containing a specific pair of domains (d, e) would interact is estimated for each pair of domains in the Bayesian manner. Finally, given a pair of host-pathogen proteins, their probability of interaction is estimated by a naive combination $(= 1 - \prod_i \prod_j (1 - P(d_i, e_j)))$ of the probabilities from each pair of domains (d_i, e_j) contained in the pair of proteins(Dyer et al., 2007).

At around the same time, Kim et al. (2007) predict *H. sapiens–H.pylori* PPIs using the PreDIN(Kim et al., 2002) and PreSPI(Han et al., 2004) algorithms, which are also based on domain information. The domain annotation used in this work is done by InterProScan as well. However, in contrast to Dyer et al. (2007), which is based on estimating the probability of an individual pair of domains being associated with protein interactions and naively combining these probabilities, PreDIN and PreSPI directly estimate the probability of domain combination pairs being associated protein interactions.

Motif-domain interaction-based approach

Some protein interactions are mediated not by interactions between domains, but by interactions between a domain in one protein and a short linear motif (SLiM) in the other protein(Edwards et al., 2007; Hugo et al., 2011). As viral pathogens typically have a compact genome, they have few domains. It is reasonable to postulate that their interaction with host proteins are likely to be mediated by Domain-SLiM interactions. For example, since HIV-1 proteins have few domains, Evans et al. (2009) predicted *H.*

sapiens–HIV-1 PPIs based on the interactions between short eukaryotic linear motifs (ELMs) and human protein counter domains (CDs).

Evans *et al.* use the ELM resource(Puntervoll et al., 2003) to determine ELMs contained in human and HIV-1 proteins and PROSITE(Hulo et al., 2008) to determine domains in human proteins. Then starting from a template human PPI (x,y) where protein x contains a ELM (E) and protein y a counter domain (CD), proteins in HIV-1 that contain the ELM (E) are predicted to form host-pathogen PPIs with the human protein y. Notably, Evans *et al.* point out that the human protein x is expected to compete with these HIV-1 proteins for interacting with y, and that this competition should be considered as another form of host-pathogen interaction.

1.2.4 Machine learning-based approach

Both supervised(Tastan et al., 2009; Dyer et al., 2011) and semi-supervised(Qi et al., 2010) learning frameworks have also been used in predicting host-pathogen PPIs. A considerable amount of interacting and non-interacting pairs are usually needed by these machine learning algorithms to produce good classifiers. For example, Tastan et al. (2009) and Qi et al. (2010) obtain curated *H. sapiens*–HIV PPIs from the 'HIV-1, human protein interaction database'(Fu et al., 2009), while Dyer et al. (2011) compile *H. sapiens*–HIV PPIs from other sources including BIND(Gilbert, 2005), DIP(Salwinski et al., 2004), IntAct(Hermjakob et al., 2004) and Reactome(Joshi-Tope et al., 2005). Supervised learning framework has first been attempted using a Random Forest (RF)(Tastan et al., 2009) classifier with 35 selected features including GO similarity, graph properties of the human interactome, ELM-ligand, gene expression, tissue feature, sequence similarity, post-translational modification similarity to neighbor, HIV-1 protein type, etc. In another work(Dyer et al., 2011), a Support Vector Machine (SVM) is used with linear kernel and features such as domain profiles, protein sequence k-mers and properties of human proteins in the human interactome.

The performance of supervised learning algorithms is limited by the availability of truly interacting proteins. However, there are a lot of protein pairs that have a known association between themselves which may not be a confirmed direct interaction(Qi et al., 2010). In order to exploit the availability of these data, Qi et al. (2010) try a semi-supervised learning approach.

The semi-supervised approach of Qi et al. (2010) use the same training data (collected by Fu et al. (2009)) as the supervised approach of Tastan et al. (2009). Tastan *et al.* use only physical PPIs with keywords "interact", "bind", etc. for training. However, Qi *et al.* use only a subset of the physical PPIs used by Tastan *et al.*. This subset consists of 158 expert-annotated *H. sapiens*–HIV PPIs and is labeled as positive training data. The remaining PPIs from Fu et al. (2009) are used as "partial positive" training data. This is because Qi *et al.* find that many of the PPIs—even those with keywords "interact", "bind", etc. —are not well agreed by experts(Qi et al., 2010). Moreover, only 18 of the 35 attributes used by Tastan *et al.* are used by Qi *et al.* Despite using fewer attributes, the separation of the PPI training data into definite known positive interactions and partial positives helps Qi *et al.* achieve a higher performance than Tastan *et al.*

An important weakness of these approaches based on machine learning is that the features used by them—e.g., the domain profile feature(Dyer et al., 2011) and the HIV-1 protein type feature(Tastan et al., 2009)—are not easy to understand, especially with respect to their biological basis. Another weakness is the limitation of training data. For example, the use of machine learning approaches in the context of host-pathogen PPI prediction has so far been applied in the *H. sapiens*–HIV system because known host-pathogen PPIs are not available in other host-pathogen systems on a sufficiently large scale.

1.3 Basic principles of host-pathogen interaction

Some basic principles derived from the analysis of experimentally verified or manually curated host-pathogen PPIs are discussed in this section. These principles either have been reported and confirmed by several works or have high potential to be applied in future works on host-pathogen interactions.

1.3.1 Topological properties of targeted host proteins

Calderwood et al. (2007) have generated 44 intra-species Epstein-Barr virus (EBV) PPIs and 173 inter-species *H. sapiens*–EBV PPIs using a stringent and systematic two-hybrid system. They observe that the degree (in the human interactome) of human proteins involved in *H. sapiens*–EBV PPIs are significantly higher than randomly selected human proteins. Thus, these targeted human proteins are enriched with hubs (i.e., proteins with high degree in the human interactome).

Moreover, Calderwood et al. (2007) also report that the minimum number of steps (in terms of PPI edges) between a targeted human protein and a reachable protein in the network is, on average, smaller than that of randomly-picked human proteins. Thus the EBV-targeted human proteins have relatively shorter paths to other proteins in the human interactome(Calderwood et al., 2007).

Dyer et al. (2008) have also analyzed the topological properties of pathogen-targeted host proteins using much larger datasets. The inter-species host-pathogen PPI and intra-species human PPI datasets studied are integrated from primary literature (Calderwood et al., 2007) and 7 databases(Gilbert, 2005; Salwinski et al., 2004; Mishra et al., 2006; Hermjakob et al., 2004; Zanzoni et al., 2002; Pagel et al., 2005; Joshi-Tope et al., 2005). This integrated host-pathogen PPI dataset contains 10,477 experimentally detected and manually curated host-pathogen PPIs, covering 190 pathogens (most of which are viruses), while the integrated human PPI dataset contains 75,457 experimentally verified PPIs(Dyer et al., 2008). The result reveals that proteins interacting

with viral and bacterial pathogen groups tend to have higher degrees (hubs), which confirms one of the observations of Calderwood et al. (2007), and higher betweenness centrality (bottlenecks).

Dyer *et al.* also analyzed the physical interaction network between human and three bacterial pathogens (*B. anthracis, F. tularensis and Y. pestis*) generated from a modified two-hybrid assay (liquid-format mating)(Dyer et al., 2010). The analyses show again pathogen preferentially interact with hubs and bottlenecks in the human interactome(Dyer et al., 2010). Zhao et al. (2011) have similarly confirmed that hubs are more likely to be targeted by viruses in studying human–virus PPIs and human signal transduction pathways.

1.3.2 Structural properties of host-pathogen PPIs

Franzosa and Xia (2011) report a significant overlap between exogenous (i.e., host-pathogen) and endogenous (i.e., within-host) interfaces of PPIs, suggesting interface mimicry as a possible pathogen strategy to evade immune system detection and to hijack host cellular machinery. The exogenous interactions represent clear cases of horizontal gene transfer between the virus and host(Franzosa and Xia, 2011). The acquisition of viral protein sequences from hosts are also observed and discussed by Rappoport and Linial (2012)

Comparing with endogenous interfaces, exogenous interfaces tend to be smaller, indicating that the viral genome is under intense selection to reduce its size compared to the host genome(Franzosa and Xia, 2011). There is a similar observation in another work(Rappoport and Linial, 2012) that viral proteins are noticeably shorter than their corresponding host counterparts, which may result from acquiring only host gene fragment, eliminating internal domain and shortening domain linkers.

Interestingly, Franzosa and Xia (2011) find that virus-targeted interfaces tend to be “date”-like. That is they are transiently used by different endogenous binding part-

ners at different times and, on average, they utilize more human binding partners than generic endogenous interfaces. This finding is supported by functional enrichment among the mimicked endogenous binding partners for the GO term "Regulation of Biological Process"(Franzosa and Xia, 2011), since proteins involved in biological regulation usually have transient binding with other proteins. This may also partially explain the topological property that targeted host proteins tend to be hubs in the host interactome(Calderwood et al., 2007), because the proteins having date-like interfaces tend to interact with many proteins and appear as hubs in intra-species PPI networks.

Lastly, an analysis of residues involved in exogenous and endogenous interfaces shows that exogenous interfaces are likely to be less conserved then endogenous interfaces(Calderwood et al., 2007).

1.4 Analysis and assessment of host-pathogen PPIs

Analysis of host-pathogen PPI datasets is essential both for developing better prediction approaches and applying the host-pathogen PPI datasets in the subsequent studies. Assessment and analysis of host-pathogen PPI datasets can be conducted directly using (i) gold standard host-pathogen PPIs or indirectly using (ii) functional information, (iii) localization information, (iv) related experimental data, (v) biological explanation of selected examples, etc.

1.4.1 Assessment based on gold standard

Known truly interacting host-pathogen PPI data (gold standard) are available for a few pathogens. The 'HIV-1, Human Protein Interaction database'(Fu et al., 2009) contains a considerable number of *H. sapiens*–HIV PPIs. A substantial number of host-pathogen PPIs (mainly *H. sapiens*–HIV PPIs) can also be found in other databases including BIND(Gilbert, 2005), DIP(Salwinski et al., 2004), IntAct(Hermjakob et al., 2004), and Reactome(Joshi-Tope et al., 2005). Therefore, in the case of *H. sapiens*–HIV PPIs, a

fairly large gold standard dataset is available. For example, the "HIV-1, Human Protein Interaction database"(Fu et al., 2009) has been used in assessing predictions based on motif-domain interaction(Evans et al., 2009). On the other hand, Davis et al. (2007) have only managed to collect 33 host-pathogen PPIs from the literature to validate their predictions for 10 pathogen species. As another example, Doolittle and Gomez (2011) have only managed to collect 3 PPIs from a public database(Dyer et al., 2008) and 20 PPIs from the literature, and only 19 among these collected PPIs are specific for the *H. sapiens*–DENV-2 system that Doolittle and Gomez (2011) have made predictions for. Although 9 of these 19 gold standard PPIs are present in the prediction results of Doolittle and Gomez (2011), the assessment has been badly hampered by the small size of the gold standard dataset.

1.4.2 Analysis and assessment based on functional information

Gene Ontology

GO terms that are significantly enriched in the host proteins predicted to be targeted by pathogens can be used to evaluate the functional relevance of the predicted host-pathogen PPIs(Davis et al., 2007). GO terms specific for human proteins involved in the immune system and for pathogen proteins involved in host-pathogen interactions can also be used to filter putative host-pathogen PPIs(Davis et al., 2007).

Several tools can analyze GO term enrichment, including GOstat(Beißbarth and Speed, 2004) used by Wuchty (2011), GO::TermFinder(Boyle et al., 2004) used by Davis et al. (2007), Ontologizer(Bauer et al., 2008) used by Tastan et al. (2009), and DAVID(Dennis Jr et al., 2003) used in many other studies(Doolittle and Gomez, 2010, 2011; Evans et al., 2009). Specifically, Wuchty (2011) analyzes the GO term enrichment of host proteins in predicted *H. sapiens–P. falciparum* PPIs and derives the 100 most enriched GO terms (in the Biological Process category) of host proteins. He finds that the pathogen may influence important signaling and regulation processes of the

host through host-pathogen PPIs(Wuchty, 2011). Tastan et al. (2009) analyze the GO term enrichment of host proteins in predicted host-pathogen PPIs; they find that 31 GO terms in the Molecular Function category (e.g., transcription regulator, ligand-dependent nuclear receptor, MHC class I receptor, and protein kinase C activities), 19 GO terms in the Biological Process category (e.g., immune system process and response to stimulus) and 14 GO terms in the Cellular Component category (e.g., membrane-enclosed lumen and plasma membrane) are significantly enriched. Enriched GO terms are identified similarly in several studies(Doolittle and Gomez, 2011, 2010) and, results show consistency with viral infection. Similarly, enriched GO terms have also been analyzed for pathogen groups(Dyer et al., 2008) and Conserved Protein Interaction Modules (CPIM)(Dyer et al., 2010) among *H. sapiens–B. anthracis, H. sapiens–F. tularensis* and *H. sapiens–Y. pestis* protein interaction networks.

Pathway data

An analysis of host-pathogen PPIs in the context of biological pathways provides a functional overview of the targeted host proteins, illuminates the mechanisms of a pathogen's obstruction on host pathways, and serves as an important assessment of predicted host-pathogen PPIs. We first discuss some results derived from an analysis of the known host-pathogen PPIs using pathway data. Then we introduce some assessment strategies of predicted host-pathogen PPIs using pathways.

Balakrishnan et al. (2009) analyze the PPI dataset from the 'HIV-1, Human Protein Interaction database'(Fu et al., 2009) in the context of human signal transduction in the Pathway Interaction Database (PID)(Schaefer et al., 2009) and Reactome(Joshi-Tope et al., 2005). They discover that a majority of human pathways can potentially be targeted by *H. sapiens*–HIV-1 PPIs. However, many alternative paths (starting and ending at the same proteins yet circumventing HIV-1 disrupted intermediate steps) to the HIV-1 targeted paths exist due to human network redundancy; and degradation

and down-regulation pathways are among the most highly targeted pathways. Singh et al. (2010) and Zhao et al. (2011) have also obtained similar results from analyzing the same pathway data: human signal transduction pathways derived from Pathway Interaction Database (PID)(Schaefer et al., 2009) and Reactome(Joshi-Tope et al., 2005) and virus–host PPI data from VirusMINT(Chatr-aryamontri et al., 2009). They find that 355 out of 671 pathways are targeted by at least one viral protein. Moreover, the majority of which (268 out of 355) are targeted by more than one viral proteins. In these 355 pathways, 413 proteins are targeted by 28 different viruses. Also, 95 of these 413 targeted host proteins are known drug targets(Singh et al., 2010; Zhao et al., 2011). However, proteins targeted by different viruses in each pathways are not necessarily the same. Zhao et al. (2011) further report that centrally-located proteins in merged networks of statistically significant pathways are hub proteins, and are more frequently targeted by viruses.

Wuchty (2011) analyzes both predicted and external (experimentally determined and structurally inferred) *H. sapiens–P. falciparum* PPIs using 184 manually curated pathways from PID(Schaefer et al., 2009). He reports that both separate and combined sets of predicted and external PPIs target proteins which have a higher degree and which appear in more pathways(Wuchty, 2011). For each pathogen protein, Wuchty (2011) identifies pathways enriched with host proteins that are targeted by this pathogen protein using Fisher's exact test. He then constructs a bipartite matrix between pathogen proteins and their corresponding enriched host signaling pathways. Observation of the matrix reveals that the pathogen has many interactions with proteins in the TNF- and NF-kappa B pathways, which indicates the pathogen's obstruction of inflammatory response(Wuchty, 2011). To evaluate host-pathogen PPIs predicted by the domain-motif interaction-based approach, KEGG pathway enrichment for HIV-1 proteins (ENV, NEF and TAT) targeted host proteins in the (experimentally verified and computationally predicted) inter-species host-pathogen PPIs are analyzed(Evans

et al., 2009). The enriched pathways include (i) immune system pathways such as T cell and B cell receptor signally pathways, apoptosis, focal adhesion, and toll-like receptor signaling pathways; (ii) disease pathways such as the colorectal cancer, leukemia and lung cancer pathways; and (iii) signal transduction processes(Evans et al., 2009).

Gene expression data

Gene expression data are another important functional information source which have been widely used in the filtering, assessment and verification of host-pathogen PPIs. Tissue-specific and infection-related gene expression data are frequently used in host-pathogen studies. A pathogen like *P. falciparum* infects different human organs at different stages of its life cycle. So the expression data of different stages of its life cycle and *H. sapiens* tissue-specific gene expression data can be used simultaneously for pruning putative *H. sapiens–P. falciparum* PPIs(Wuchty, 2011; Krishnadev and Srinivasan, 2008). For example, *P. falciparum* invades *H. sapiens* liver tissue during the sporozoit stage. The predicted host-pathogen PPIs are thus more likely to be real, if the corresponding human proteins are known to express in liver tissue and the corresponding pathogen proteins are known to express in the sporozoit stage. This filtering strategy has been adopted by several studies(Wuchty, 2011; Krishnadev and Srinivasan, 2008). For the *H. sapiens–M. tuberculosis* system, human proteins expressed in lung tissue or bronchial epithelial cells and pathogen proteins upregulated in granuloma, pericavity, or distal infection sites can be used for filtering purposes(Davis et al., 2007). Moreover, pathogen genes involved in *M. tuberculosis* infections(Sassetti and Rubin, 2003; Rachman et al., 2006), and human genes involved in *M. tuberculosis, L. major, T. gondii* infections(Chaussabel et al., 2003) can be compared with the pathogen and host proteins in predicted *H. sapiens–M. tuberculosis* PPIs as a useful assessment(Davis et al., 2007).

RNA interference data

RNA interference (RNAi) is a natural process to specifically and selectively inhibit a targeted gene expression. Small interfering RNA (siRNA), short hairpin RNA (shRNA) and bi-functional shRNA are often used to mediate the RNAi effect. Some human proteins, when being silenced by genome-wide RNAi experiments, are found not lethal to human cells but essential for HIV replication. Those human proteins may have high likelihood of interacting with HIV. Therefore, comparing the set of host proteins in predicted host-pathogen PPIs and the set of host proteins identified by RNAi experiments can be used as an assessment. We briefly list some examples below.

Several studies show that knocking down some host proteins by siRNA(König et al., 2008; Brass et al., 2008; Zhou et al., 2008) or shRNAs(Yeung et al., 2009), can impair HIV-1 infection or replication. Thus, those host proteins are essential for HIV-1 infection or replication. Therefore, they have higher possibility to interact HIV-1 proteins. This has been used as a filtering criterion(Doolittle and Gomez, 2010) and assessment data(Tastan et al., 2009; Qi et al., 2010; Dyer et al., 2011; Evans et al., 2009) in several studies.

Three works(Tastan et al., 2009; Qi et al., 2010; Dyer et al., 2011) based on the machine learning approach for predicting *H. sapiens*–HIV PPIs use a siRNA dataset(Brass et al., 2008) to assess their prediction results. The assessment is conducted by examining the overlap between the human proteins targeted by the predicted PPIs and the proteins in the siRNA dataset(Brass et al., 2008). Besides Qi et al. (2010) also combine four RNAi datasets(König et al., 2008; Brass et al., 2008; Zhou et al., 2008; Yeung et al., 2009) and conduct additional assessment in a similar way.

A five-way comparison has been conducted by Evans et al. (2009) on five HIV-1 targeted human protein datasets—viz., (i) the human protein dataset targeted by PPIs predicted using the motif-domain interaction-based approach(Evans et al., 2009); (ii) human protein dataset targeted by gold standard PPIs from the ‘HIV-1,Human

Protein Interaction database'(Fu et al., 2009); and (iii) human protein datasets from three genome-wide RNAi experiments(König et al., 2008; Brass et al., 2008; Zhou et al., 2008). Results show that genome-wide RNAi experiments match each other better than the interaction studies(Evans et al., 2009). The matches between protein dataset (i) and the other four protein sets are significant, but discrepancies are still observed(Evans et al., 2009).

For the *H. sapiens*–DENV system, host protein datasets from two siRNA experiments in DENV infection(Sessions et al., 2009; Krishnan et al., 2008) are available. They have also been used to refine *H. sapiens*–DENV PPI prediction result(Doolittle and Gomez, 2011).

1.4.3 Pruning based on localization information

Localization information of pathogen and host proteins may relate to the possibility of their interactions. For extracellular pathogens, their extracellular or secretion proteins may have higher chance of interacting with host surface proteins rather than host nuclear proteins. For intracellular pathogens like viruses, co-localization of host and pathogen proteins may be one of the prerequisites for protein interactions. Several studies use these information to filter prediction results.

Sub-cellular localization of host and pathogen proteins

Since pathogen extracellular and secretion proteins, and proteins with translocational signals are more likely to interact with host extracellular or membrane proteins, such sub-cellular localization information are often used in pruning of predicted host-pathogen PPIs(Lee et al., 2008; Krishnadev and Srinivasan, 2008; Tyagi et al., 2009; Krishnadev and Srinivasan, 2011; Wuchty, 2011). In connection with this, several tools are used in homology-based approaches(Krishnadev and Srinivasan, 2008; Tyagi et al., 2009; Krishnadev and Srinivasan, 2011) to predict protein sub-cellular localization.

Co-localization of host and pathogen proteins

As obligate intracellular pathogens, viruses do not have cellular structure or their own metabolism, and are solely dependent on the host cell. Therefore, a viral protein and its host protein interaction targets are more likely to be co-localized. Several studies use this basic assumption to assess or filter predicted *H. sapiens*–HIV PPIs(Doolittle and Gomez, 2010) and *H. sapiens*-DENV PPIs(Doolittle and Gomez, 2011). Similar information is also used as one of the selected features for classifiers in approaches based on machine learning for predicting *H. sapiens*–HIV PPIs(Qi et al., 2010; Tastan et al., 2009). The co-localization information of two proteins can be revealed through their shared GO terms in the Cellular Compartment category.

1.4.4 Biological explanation of selected examples

An analysis of a specific PPI by explaining the underlining biological functions is not an effective assessment of predicted host-pathogen PPIs, because such an analysis can cover only a small number of PPIs. However, it may facilitate a better understanding of that putative PPI, and therefore promote subsequent experimental verification of that prediction. Explanation of the biological basis of some example PPIs from the whole dataset can be found in many studies(Krishnadev and Srinivasan, 2008; Tyagi et al., 2009; Krishnadev and Srinivasan, 2011; Dyer et al., 2011; Davis et al., 2007; Doolittle and Gomez, 2011, 2010). Some of the specific examples may have literature or experimental supports, some lack direct literature support but have some indirect supports including structural information, homology to template PPIs, evidence from related experiments (gene expression and RNAi experiment data), etc. Explanation and identification of validated predictions also enhance the impact of the prediction methods; and this approach has been used in many studies(Davis et al., 2007; Dyer et al., 2011; Krishnadev and Srinivasan, 2011). For example, Dyer et al. (2011) discuss in detail the predicted *H. sapiens*–HIV PPIs involving the HIV Dependency Factors(Brass et al.,

2008) that have support in the literature(Dyer et al., 2011). To some extent, explanation of indirect evidence and clues enhances validity of the selected parts of the prediction results(Krishnadev and Srinivasan, 2011; Tyagi et al., 2009; Davis et al., 2007; Doolittle and Gomez, 2011, 2010). Predicted PPIs both with and without experimental verifications, and PPIs involving hypothetical proteins are discussed and explained in Krishnadev and Srinivasan (2011). In another work, Tyagi et al. (2009) also explain some examples of predicted *H. sapiens–H. pylori* PPIs through the structural point of view, and discuss examples of PPIs involving membrane proteins, secreted proteins and hypothetical proteins.

1.4.5 Assessment through related experimental data

Some related experimental data turn out to be useful for assessing the targeted host proteins in host-pathogen PPIs. For example, during budding, host proteins may be incorporated into virion(Chertova et al., 2006). Although some host proteins may be taken up by a budding virus accidentally, others are known to play crucial roles in viral life cycle and host-pathogen interaction. A dataset(Chertova et al., 2006) on human protein presents in virion has also been used to filter predicted *H. sapiens*–HIV-1 PPIs(Doolittle and Gomez, 2010).

Qi et al. (2010) and Tastan et al. (2009) use a human protein set hijacked by HIV-1 into its virion(Ott, 2008) to assess their predicted *H. sapiens*–HIV-1 PPIs. Specifically they examine the overlap between targeted human proteins in the predicted PPIs and the 314 human proteins in virion(Ott, 2008). A large overlap suggests a satisfactory performance of the prediction approach(Tastan et al., 2009; Qi et al., 2010).

1.5 Host-pathogen interaction data collection and integration

The rapid progress on the host-pathogen interaction studies is supported by many collection, dissemination, integration, analysis and visualization tools. Host-pathogen interaction databases, can be divided into two categories: (i) collection and curation databases; (ii) integration and analysis databases. There is no clear dividing line for the two categories. This categorization is mostly for convenience of discussion.

1.5.1 Host-pathogen interaction data collection techniques

Text mining is frequently used for extracting PPI data from literature. This is very useful in facilitating the manual curation of host-pathogen interaction data from publications. For example, VirusMINT(Chatr-aryamontri et al., 2009) relies on a simple text mining approach based on a context-free grammar that identifies sentences containing interaction information to select relevant articles. VirHostNet(Navratil et al., 2009) also uses a text mining approach to prioritize papers for manual curation, where the text mining pipeline is applied to extract keywords related to both virus and experimental procedures.

Moreover, text mining techniques have been applied to specifically extract host-pathogen PPIs from biomedical literature with considerable accuracy(Thieu et al., 2012). Feature-based and language-based approaches are introduced and compared by Thieu et al. (2012). Both methods can automatically detect host-pathogen interaction data and extract information about organisms and proteins involved in the interactions(Thieu et al., 2012). The feature-based method uses SVM trained on features derived from the individual sentences, including names of the organisms and corresponding proteins or genes, keywords describing host-pathogen interaction-specific information, general PPI information, experimental methods, and other statistical in-

formation(Thieu et al., 2012). The language-based method uses a link grammar parser combined with semantic patterns derived from training examples(Thieu et al., 2012).

1.5.2 Host-pathogen interaction collection and curation databases

Host-pathogen interaction collection and curation databases are those dedicated to collect and curate host-pathogen interaction from literature or from experimental data. These databases may have imported some parts of data from other databases but at least contain some data derived from their own collection or curation. Collection and curation databases serve primarily as "data source", and generally provide only simple tools for searching, visualization or analysis. They are often used as the data source for host-pathogen interaction studies or are imported by integration and analysis databases (to be discussed in the next section). In this section we list some representative databases of this category.

PHI-base is a database created to catalog experimentally verified pathogenicity, virulence and effector genes of fungal and Oomycet pathogens(Winnenburg et al., 2006). After its update, PHI-base also covers bacterial pathogens. The pathogens covered by PHI-base infect a wide range of hosts(Winnenburg et al., 2008).

The 'HIV-1, Human Protein Interaction database' at NCBI aims at cataloging all interactions between HIV-1 and human proteins published in the peer-reviewed literature(Fu et al., 2009). Basic search and visualization tools are also provided. It is very popular among the AIDS research community. It is well known for its intensive long-term curation effort. The *H. sapiens*–HIV-1 interaction data included in this database cover both direct and indirect interactions; brief description and PubMed IDs are also provided for each entry. Its *H. sapiens*–HIV-1 PPI data have been used in several studies(Tastan et al., 2009; Qi et al., 2010; Evans et al., 2009) and imported as source data by some databases(Chatr-aryamontri et al., 2009).

The VirusMINT database aims at collecting all interactions between viral and hu-

man proteins reported in the literature(Chatr-aryamontri et al., 2009). It covers more than 110 different viral strains(Chatr-aryamontri et al., 2009). The curation effort has focused mainly on viruses known to be associated with infectious diseases and oncogenesis in humans(Chatr-aryamontri et al., 2009). VirusMINT derives its host-virus PPI data from two sources. The first source is from databases of literature-curated PPIs like IntAct(Hermjakob et al., 2004), MINT(Zanzoni et al., 2002), and 'HIV-1,Human Protein Interaction database'(Fu et al., 2009). Host-virus PPI data are uploaded from IntAct and MINT directly without further curation. Only a subset of 'HIV-1, Human Protein Interaction database' is imported, which pertains to enzymatic reactions, physical associations and co-localization. The second source are PPIs manually curated from literature; the PPIs are first uploaded to MINT and then re-imported into VirusMINT(Chatr-aryamontri et al., 2009). The literature curation is facilitated by simple text ming techniques in selecting relevant articles. MINT(Zanzoni et al., 2002) and VirusMINT are both curated by MINT curators and uploaded first to MINT then to VirusMINT. Much of the PPI data in VirusMINT are the same as in MINT. VirusMINT also provides searching and visualization functions.

VirHostNet (Virus-Host Network) is a management and analysis database of integrated virus-virus, virus-host and host-host interaction networks and their functional annotations(Navratil et al., 2009).The interaction data are reconstructed from public databases and, for virus-virus and virus-host interactions, are also supplemented by original literature-curated dataset.

A simple text mining strategy has been adopted for prioritizing articles for literature curation. Virus-virus and virus-host interactions data from public databases are also carefully inspected before importing into VirHostNet(Navratil et al., 2009). Search and visualization functions are supported in this database.

The databases mentioned below are mostly well known for their intra-species PPI datasets. However, their curation and collection have also been extended to inter-

species host-pathogen PPIs. IntAct is an open-source, open-data molecular interaction database(Kerrien et al., 2012). Both intra and inter-species PPI data are collected in this database either from the literature or from direct data depositions. For each PPI entry a brief description, experimental method and literature citation are included. Several integration databases(Chatr-aryamontri et al., 2009; Driscoll et al., 2009; Ranjit and Bindu, 2010) import host-pathogen PPI data from IntAct. It is well known for its intensive curation and quality control process. BioGRID (Biological General Repository for Interaction Datasets) archives and disseminates genetic and protein interaction data(Stark et al., 2011). BioGRID interaction data are curated from both high-throughput experiments and individual focused studies. Most of the interaction data are intra-species PPIs, but some host-pathogen PPIs are included. DIP (Database of Interacting Proteins) aims to integrate the diverse experimental evidences on PPIs into the database(Salwinski et al., 2004). It is another well-known intra-species PPI integration database. It also collects host-pathogen PPI data. Reactome is a curated, peer-reviewed knowledgebase of biological pathways(Joshi-Tope et al., 2005). It curates both intra- and inter-species data. Curated host-pathogen PPI data are also available in Reactome(Joshi-Tope et al., 2005). BIND (Biomolecular interaction network database) archives biomolecular interaction, complex and pathway information, and is a major source of curated biomolecular interactions(Gilbert, 2005). It has not been maintained for the last few years, until a recent update and conversion of the BIND data to a standard format (Proteomics Standard Initiative-Molecular Interaction 2.5)(Isserlin et al., 2011). Its main interaction data are intra-species PPIs, but also contains some host-pathogen PPI data.

1.5.3 Host-pathogen interaction integration and analysis databases

Host-pathogen interaction integration and analysis databases mainly integrate host-pathogen interaction data from other source databases. While they usually do not

have their own intensive curation process, some of them provide powerful analysis and visualization functions. The integrated data can be more than just host-pathogen PPI data, like gene expression data related to infection, disease outbreak information, pathogen proteomics data, protein functional data, protein complex data, etc. In this section, representative integration and analysis databases are briefly introduced.

APID (Agile Protein Interaction Data Analyzer) provides an open-access framework where all known experimentally validated protein-protein interactions (BIND, BioGRID, DIP, HPRD, IntAct and MINT) are unified in it(Prieto and De Las Rivas, 2006). iRefIndex(Razick et al., 2008) provides an index of PPIs from BIND, BioGrid, DIP, HPRD, IntAct, MINT, MPact(Güldener et al., 2006), MIPS(Pagel et al., 2005) and OPHID(Brown and Jurisica, 2005). iRefWeb(Turner et al., 2010) provides a searchable web interface to the iRefIndex. Both APID and iRefIndex (iRefWeb) are general PPI integration databases, unlike the following databases which are dedicated to host-pathogen interaction data integration and analysis. They include host-pathogen PPIs just because their source databases contain some host-pathogen PPI data.

PHIDIAS (Pathogen-Host Interaction Data Integration and Analysis System) includes six components (PGBrowser, Pacodom, BLAST searches, Phinfo, Phigen and Phinet) for searching, comparing, and analyzing integrated genome sequences, conserved domains, host-pathogen interaction data and gene expression data related to host-pathogen interactions(Xiang et al., 2007).

HPIDB is a host-pathogen PPI database which integrates experimental PPIs from several public databases (BIND, REACTOME, MINT, IntAct, PIG)(Ranjit and Bindu, 2010). Some of the HPIDB sources may have content overlap with each other, since PIG(Driscoll et al., 2009) also integrates data from BIND, REACTOME, and MINT. Different from PIG—which only considers one host, *H. sapiens*—HPIDB also takes other hosts into account.

GPS-Prot is an integration and visualization database that currently focuses on

H. sapiens–HIV interactions(Fahey et al., 2011). It allows for integration of different HIV interaction data types(Fahey et al., 2011). Human PPI data are imported from the following six databases, MINT, IntAct, DIP, MIPS, BioGRID and HPRD. *H. sapiens*–HIV PPI data are import from VirusMINT(Chatr-aryamontri et al., 2009). The GPS-Prot can group proteins into functional modules or protein complexes, generating intuitive network representations. It allows for the uploading of user-generated data(Fahey et al., 2011).

RCBPR (Resource Center for Biodefense Proteomics Research) is a bioinformatics framework employing a protein-centric approach to integrate and the collect large and heterogeneous data(McGarvey et al., 2009). It is no longer functional and the collected data have been transferred to the Pathogen Portal (http://www.pathogenportal.org).

PIG (Pathogen Interaction Gateway)(Driscoll et al., 2009) is created by integrating host-pathogen PPI data from a number of public resources, including BIND, REACTOME, MINT, MIPS, HPRD, DIP, and MvirDB(Zhou et al., 2007). Now PIG has become part of the PATRIC(Gillespie et al., 2011) database; but only the bacterial pathogen data in PIG have been merged into PATRIC.

Disease View is a host-pathogen data integration and visualization resource that enables access, analysis, and integration of diverse data sources, including host, pathogen, host-pathogen interactions, and disease outbreak. It provides a mechanism for infectious disease-centric data analysis and visualization. The infections diseases covered by Disease View come with related information like the corresponding pathogen that causes the infectious diseases, the associated pathogen virulence genes and the genetic and chemical evidences for the human genes that are associated with the diseases(Driscoll et al., 2011). It is implemented as a component of PATRIC(Gillespie et al., 2011).

PATRIC (the Pathosystems Resource Integration Center) is a comprehensive genomics-centric relational database for infectious-disease research(Gillespie et al., 2011). Com-

prehensive bacterial genomics data, associated data relevant to genomic analysis, and analysis tools and platforms have been provided in this database. Its resources can be divided into two categories, (i) organisms, genomes, and comparative genomics; (ii) recurrent integration of community-derived associated data.

1.5.4 Host-pathogen interaction integration and analysis software

Not only databases but also standalone software tools are available for host-pathogen interaction studies.

Conventional complex network analysis and visualization software platforms like Cytoscape(Smoot et al., 2011) continue to be very popular in host-pathogen interaction studies. Cytoscape has been used for visualization of host-pathogen PPI networks in several works(Dyer et al., 2007, 2008). Software that are specifically designed for host-pathogen interaction studies have also been developed. For example, BiologicalNetworks is a system that enables the integration of multi-scale data for host-pathogen studies(Sergey et al., 2011). It can integrate diverse experimental data types, including molecular interactions, phylogenetic classifications, genomic sequences, protein structure information, gene expression, pathway and virulence data for host-pathogen studies(Sergey et al., 2011). It provides several useful functions including, analyzing sub-networks, building host-pathogen interaction networks, studying individual genes, identifying potential drug targets, adding phylogeography, integrating user data, etc(Sergey et al., 2011). This system is available through a standalone Java application (BiologicalNetworks), which provides complex data analysis capabilities, and a web interface (http://flu.sdsc.edu) for quick search of phylogenetic relations among sequenced strains.

1.6 Discussion

1.6.1 Contributions and limitations of current host-pathogen interaction study approaches

The current host-pathogen interaction studies described in this Chapter are indispensable stepping stones for the future progress in this field. Nevertheless, several limitations are also noticeable.

Contributions of current host-pathogen interaction studies

Usually host-pathogen PPIs prediction followed by analyses and assessment would produce enriched datasets which are useful for the experimental testing and verification. This could save a lot of wet lab experimental effort. The prediction and verification approaches discussed in these pioneering works pave the way for future development of host-pathogen interaction studies as they provide insights for improvements and basis for comparison.

Limitations of current host-pathogen interaction prediction approaches

It is not uncommon that different prediction approaches yield very different prediction results, even in the same host-pathogen system, as revealed by the comparison among different *H. sapiens*–HIV PPI datasets generated from different prediction approaches(Doolittle and Gomez, 2010).

It has not escaped our notice that some publications repeatedly report almost the same prediction method whose performance and predicted results have not been rigorously assessed. Sometimes even the source data (like template PPI data) are the same, yet only applied to different host-pathogen systems(Krishnadev and Srinivasan, 2008; Tyagi et al., 2009; Krishnadev and Srinivasan, 2011). Therefore the contribution of these publications may be relatively limited.

Limited by the current understanding of host-pathogen protein interaction, the prediction approaches may not resemble the real biological scenario. For example, although the approach based on motif-domain interaction(Evans et al., 2009) has achieved good performance, Evans *et al.* has also mentioned the mismatches between predicted result and gold standard may be caused by the fact that real mechanisms of host-pathogen PPIs are more complicated than the assumption (that host-pathogen PPIs are mediated by ELMs-CDs interactions) in this study.

Limitations of current host-pathogen interaction verification approaches

Due to the limitation of current known gold-standard host-pathogen PPI data and limited understanding of the host-pathogen interactions, most current assessments are rather "indirect" approaches.

Some verifications may not have a strong logical or biological basis. For example, Dyer et al. (2007) assessed predicted *H. sapiens-P. falciparum* PPIs by examining whether the pairs of human proteins predicted to interact with the same pathogen proteins are close to each other in the human PPI network. This assessment through distance in triplets may not have biological or experimental basis. However, based on the observed topological properties discussed in the Section 1.3, "Basic principles of host-pathogen interaction", calculating whether the human proteins targeted by predicted PPIs have shorter paths to other reachable proteins in the human interactome, would serve as a possible assessment. Dyer et al. (2007) also analyze the gene expression profile of pathogen protein pairs interacting with the same host proteins; they report that those pathogen protein pairs exhibit correlated gene expression profile, and also the same for host protein pairs interacting with same pathogen proteins. While gene expression profile can be reasonably used in assessing *M. tuberculosis* H37Rv intra-species PPI datasets as done by Zhou and Wong (2011), it may lack biological basis in assessing inter-species host-pathogen PPI dataset through gene expression in the form

of triplets as conducted by (Dyer et al., 2007).

Explanation on selected examples of predicted results, neither reflects the quality of the whole predicted results nor the performance of prediction approaches. For example, biological explanation for selected examples should not be used as the only assessment of a few predicted results, as what we observed in several studies(Lee et al., 2008; Krishnadev and Srinivasan, 2008; Tyagi et al., 2009; Krishnadev and Srinivasan, 2011)—the qualities of those prediction results are still largely in doubt.

1.6.2 Contributions and limitations of current host-pathogen interaction databases

Current host-pathogen interaction databases contribute a lot to host-pathogen interaction studies in the form of collecting and integrating valuable host-pathogen interactions and providing powerful analysis tools. Yet some possible limitations also exist.

The host-pathogen interaction databases greatly facilitate host-pathogen interaction studies in collecting and integrating valuable interaction and related genomic and experimental data scattered in primary literature. Without these databases, many of the studies described in this Chapter would be impossible or at least would take much longer time and more effort in collecting the source data. Moreover, these databases provide the platforms for accessing and sharing of host-pathogen interaction data, which in turn facilitate research in this field. Many databases not only enable convenient data access and integration of related host and pathogen data, but also provide powerful analysis tools which significantly increase the efficiency of host-pathogen interactions analysis.

Some databases lack long-term support and are no longer in function, like RCBPR (McGarvey et al., 2009; Zhang et al., 2008). And there are some information loss in the merging of the one database into another, like PIG(Driscoll et al., 2009), where only its bacterial pathogen data have been moved into PATRIC. Some databases, although still

Limitations	databases
Lack long-term support(no longer in function)	RCBPR
Information loss in the merging to another database	PIG
Lack necessary updates(still in operation)	ViursMINT, HPIDB

Table 1.1: Summary of limitations of current host-pathogen interaction databases

in operation, lack necessary updates, like ViursMINT(Chatr-aryamontri et al., 2009) and HPIDB(Ranjit and Bindu, 2010); refer Table 1.1.

1.6.3 Literature-curated host-pathogen interaction data

The literature-curated interaction data from the databases discussed above are often used as gold standard in studies on host-pathogen interactions. However, a study(Cusick et al., 2008) on intra-species PPI datasets shows that literature-curated PPI data may not be as accurate as people usually have assumed. Therefore, those manually curated host-pathogen PPI data should be used with caution. For example, the 'HIV-1, Human Protein Interaction database' at NCBI(Fu et al., 2009) has been divided into "positively labeled" and "partially labeled" data in Qi et al. (2010), and VirusMINT only imports a portion of the PPI data from it.

1.6.4 Future development of host-pathogen interaction studies

Fundamental progress in host-pathogen interaction studies will be achieved in the future, due to better source data, improved investigation approaches and tools, and these will lead to deeper understanding of host-pathogen interaction.

It is reasonable to expect that high-quality source data will become increasingly available. More genomic and proteomic data will emerge. As a result, more accurate orthologs can be identified between less-known pathogens and well-studied organisms. This will enable the application of homology-based approach to many understudied host-pathogen systems. Better annotation of known motifs and counter domains will result in enhanced performance of domain-motif interaction-based predic-

tion approaches(Evans et al., 2009). With more protein structures being resolved, structure-based approach will have higher-quality structural template and with much larger coverage. With more high-resolution Structural Interaction Network (SIN) being provided to analysis, more fundamental interaction mechanisms will come to light. Abundant and accurate functional information—including, GO annotation, gene expression, RNA interference, and pathway data—will largely improve the performance of current analysis approaches. More reliable PPI data (both intra- and inter-species) will provide sufficient high-quality templates for homology-based approaches, and also larger as well as more accurate training and testing data for machine learning-based approaches. The lack of gold-standard host-pathogen PPI data will also be alleviated in the future. As a result more direct and effective verification approaches will be available for many host-pathogen systems. With better source data from a variety of aspects, the prediction approaches that can integrate different types of data (e.g., machine learning-based approach) into their prediction will have good potential.

More effective host-pathogen interaction prediction algorithms will be proposed in the future. For example, the core algorithm used by Dyer *et al.* is an association method proposed in 2001. However, several other algorithms with enhanced performance are available now, including the association numerical method (ASNM)(Hayashida et al., 2004) in 2004 and the association probabilistic method (APM)(Chen et al., 2006) in 2006. Using ASNM or APM in predicting host-pathogen PPIs may improve prediction performance. Recently, Itzhaki et al. (2010) propose the concept of "preferential use of protein domain pairs as interaction mediators" may also introduce new idea to DDI-based prediction algorithms. More accurate prediction and more effective verification approaches on a better understanding of host-pathogen interactions will come out.

All these will help the community to achieve the ultimate goal of better prevention and treatment of infectious diseases.

1.7 Objective of this dissertation

This dissertation aims at conducting a systematic study on host–pathogen PPIs based on the *H. sapiens–M. tuberculosis* H37Rv model system. The systematic study consists of three parts:(i)identifying reliable pathogen PPI dataset; (ii)constructing comprehensive functional analysis tool; (iii) developing more accurate prediction approaches and analyzing the predicted inter-species host–pathogen PPIs.

Intra-species PPI datasets are crucial for understanding the functional role of the proteins that are involved in host–pathogen PPIs. And intra-species pathogen PPIs can be used as training data for the prediction of host–pathogen PPIs as in Dyer et al. (2007). Like most pathogens, large-scale high-quality intra-species pathogen PPIs of *M. tuberculosis* H37Rv are not readily available. If we want to conduct a systematic study on *H. sapiens–M. tuberculosis* H37Rv PPIs, a reliable *M. tuberculosis* H37Rv intra-species PPI dataset has to be identified. Therefore, in the first part of our systematic study, we work towards building the foundation of *H. sapiens–M. tuberculosis* H37Rv PPI study by identifying a reliable *M. tuberculosis* H37Rv intra-species PPI dataset.

The lack of high-quality large-scale inter-species PPIs is a problem common to most host–pathogen systems, including *H. sapiens–M. tuberculosis* H37Rv. This results in a lack of gold standard for assessing the predicted host–pathogen PPIs. Thus, functional analysis based on pathway data are frequently used to assess the predicted host–pathogen PPIs. But due to several major limitations of current pathway databases, the effectiveness of pathway data for analysis and assessment of predicted host–pathogen PPIs has been seriously reduced. Therefore, the second part of our systematic study is to create our own analysis tool—IntPath. IntPath enables comprehensive functional analysis based on integrated pathway data for both host and pathogen.

Predicting host–pathogen PPIs is one of the important topics in host–pathogen interaction studies, as the experimental host–pathogen PPIs are usually very scarce. Many prediction approaches have been proposed in the past few years as discussed in

Section 1.2. However the performance of most prediction approaches are rather limited and their accuracy is largely unknown. Therefore, for the third part of our systematic investigation, we aim at developing more accurate prediction approaches than existing approaches. We tackle the challenge in two ways, by refining the homology-based prediction approach and the DDI-based prediction approach.

We compare our refined approaches to the corresponding conventional DDI-based and homology-based approaches in terms of cellular compartment distribution, disease gene list enrichment, pathway enrichment and functional category enrichment. The analysis results support the validity of our prediction result and clearly show that our our refinements lead to better performance in predicting *H. sapiens–M. tuberculosis* H37Rv PPIs.

This dissertation presents our systematic investigation on host–pathogen PPIs through three major parts as discussed above. The tools, methods, prediction results and observed properties described in this dissertation can be an important stepping stone for the future host–pathogen interaction studies.

1.8 Declaration

This dissertation is based on the following material:

- "Progress in computational studies of host-pathogen interactions", H. Zhou, J. Jin and L. Wong. *J. Bioinform. Comput. Biol.*, 11(2):1230001, April 2013.
- "Comparative analysis and assessment of *M. tuberculosis* H37Rv protein-protein interaction datasets", H. Zhou and L. Wong. *BMC Genomics*, 12(Suppl 3):S20, November 2011.
- "IntPath—an integrated pathway gene relationship database for model organisms and important pathogens", H. Zhou, J. Jin, H. Zhang, Y. Bo, M. Wozniak and L. Wong. *BMC Systems Bio.*, 6(Suppl 2):S2, December 2012.

- "Stringent DDI-based prediction of *H. sapiens–M. tuberculosis* H37Rv proteint-protein interactions", H. Zhou, J. Rezaie, W. Hugo, J. Jin, C. H. Yong, M. Wozniak and L. Wong. *BMC Systems Bio.* Accepted, September 2013.

- "Accurate homology-based prediction of *H. sapiens–M. tuberculosis* H37Rv proteint-protein interactions", H. Zhou, N. Nguyen, J. Jin, L. Zhao, M. Fan and L. Wong. *BMC Systems Bio.*, submitted, May 2013.

Chapter 2

Analysis of *M. tuberculosis* H37Rv PPI Datasets

M. tuberculosis protein-protein interactions (PPIs) data are crucial for the study of host-pathogen interaction. However, the quality of the well known *M. tuberculosis* PPI datasets is unclear. This hampers the effectiveness of research works that rely on these PPI datasets. Here, we analyze the two well known *M. tuberculosis* H37Rv PPI datasets. The first dataset is the high-throughput B2H PPI dataset from a recent paper in *Journal of Proteome Research* (Wang et al., 2010). The second dataset is from the STRING database, version 8.3, which is comprised entirely of H37Rv PPIs predicted using various methods.

We find that these two datasets have a surprisingly low level of agreement. To test the quality of these two datasets, we evaluate them based on correlated gene expression profiles, and coherent informative GO term annotations. Next, to test the possibility that the H37Rv STRING PPIs are not purely direct physical interactions, we compare B2H PPIs and predicted PPIs in STRING for protein pairs that catalyze adjacent steps in enzymatic reactions. If two proteins in a pair catalyze adjacent steps in enzymatic reactions, these two proteins are very likely have close functional association; however,

they may not have physical interactions.

The results show that the H37Rv B2H PPI dataset is of low quality. It should not be used as the gold standard to assess the quality of other (possibly predicted) H37Rv PPI datasets. The H37Rv STRING PPI dataset is also of low quality; nevertheless, a subset consisting of STRING PPIs with score $\geq$770 has satisfactory quality. However, these STRING "PPIs" should be interpreted as functional associations, which include a substantial portion of indirect protein interactions, rather than direct physical interactions. These two factors cause the strikingly low similarities between these two main H37Rv PPI datasets. The results and conclusions from this comparative analysis provide valuable guidance in using these *M. tuberculosis* H37Rv PPI datasets in subsequent studies for a wide range of purposes.

2.1 Background

M. tuberculosis H37Rv protein-protein interaction (PPI) data has become an important reference for the host-pathogen interaction studies. However, *M. tuberculosis* H37Rv PPI data is far from complete and accurate. Hitherto predicted *M. tuberculosis* H37Rv PPIs in the STRING database (version 8.3 contains 248,574 PPIs covering 3,965 proteins)(Szklarczyk et al., 2011) have seen the most frequent use because large-scale experimental PPI datasets have not been available until recently.

The first large-scale proteome-wide PPI dataset of H37Rv was produced in 2010 using a high-throughput bacterial two-hybrid (B2H) approach (Wang et al., 2010); it comprises 8,042 PPIs covering 2,907 proteins. No doubt in the foreseeable future, increasingly more studies on *M. tuberculosis* will base on both of these datasets.

There is an extremely low overlap of just 276 protein-protein interactions shared between the 8,042 H37Rv PPIs (covering 2,907 proteins) in the B2H dataset and the 248,574 predicted H37Rv PPIs in STRING(covering 3,965 proteins). In fact a hypergeometric test—drawing 8,042 pairs from the background of 7,858,630 (= 3,965 choose

2) possible pairs, and testing for an intersection of size at least 276 with the 248,574 predicted H37Rv PPIs in STRING—gives a p-value of 0.9, showing that the overlap of the STRING and B2H PPI datasets is no better than random. It is the objective of this Chapter to investigate the cause of this unexpectedly low overlap. We hypothesize that this low overlap between the two datasets may be due to (i) the B2H dataset is poor in quality, (ii) the STRING dataset is poor in quality, and/or (iii) the STRING dataset does not correspond to direct physical protein-protein interactions.

In order to test the quality of these two *M. tuberculosis* H37Rv PPI datasets, we evaluate them based on correlated gene expression profiles, coherent informative GO term annotations, and conservation in other organisms. Two proteins that interact are expected to be expressed at the same time, and be located or transported to the same place. Thus their underlying genes are likely to exhibit correlated expression profiles. Two proteins interact to effect a biological process or molecular function; thus they are expected to be annotated to some GO terms in common or GO terms that are closely related. Many protein-protein interactions are expected to be conserved across several organisms that have common ancestry; thus real protein interactions are likely to coincide with interologs from such organisms.

These assessments indicate H37Rv B2H PPIs agree less well with correlated gene expression profiles, coherent informative GO term annotations, and conservation in other organisms than H37Rv STRING PPIs (with score $\geq$ 770) . This suggests that PPIs in the H37Rv B2H dataset may contain a high level of noise (false positives).

As mentioned earlier, the H37Rv STRING PPI dataset (with score $\geq$ 770) , which is comprised entirely of PPIs predicted using a variety of methods, shows good agreement with correlated gene expression profiles, coherent informative GO term annotations, and conservation in other organisms. However, protein pairs that are functionally linked are also expected to agree well with correlated gene expression profiles, coherent informative GO term annotations, and conservation in other organisms, even though

many of these protein pairs do not have direct physical interactions. In order to test whether the predicted PPIs in STRING correspond to direct physical protein-protein interactions, we should analyze the similarity between these predicted PPIs with several distinct types of protein pairs such as experimental PPIs obtained from two-hybrid assays, protein pairs that belong to the same protein complexes, and protein pairs that catalyze adjacent steps in enzymatic pathways. As these types of additional information are not available for *M. tuberculosis*, we turn to the model organism *S. cerevisiae* where more comprehensive information is available. We extract from STRING an unbiased representative *S. cerevisiae* PPI subset (which we denote "predicted functional associations dataset") that are predicted using similar methods as the H37Rv STRING PPI dataset. For the three different types of protein pairs, we use the following gold standard: (i) *S. cerevisiae* two-hybrid PPI dataset from Yu et al. (2008), (ii) all protein pairs found in the same *S. cerevisiae* protein complexes from Wodak Lab (Pu et al., 2009), and (iii) protein/gene pairs that catalyze/form successive reaction steps in biological pathways from KEGG (Ogata et al., 1999), Wikipathways (Pico et al., 2008; Kelder et al., 2009) and Biocyc (Karp et al., 2005).

This analysis indicates that the predicted *S. cerevisiae* STRING PPIs show higher similarities with protein pairs in the same protein complexes and protein/gene pairs that catalyze/form adjacent reaction steps in biological pathways than with PPIs reported in two-hybrid assays. Therefore, the predicted *S. cerevisiae* STRING PPIs are mostly not direct physical protein-protein interactions. As the H37Rv STRING PPIs are predicted using similar methods, in turn, they are also unlikely to correspond to direct physical interactions. Nonetheless, their relatively good agreement with correlated gene expression profiles, coherent informative GO term annotations, and conservation in other organisms suggests that the H37Rv STRING PPIs (at score $\geq$ 770) are proteins that are functionally linked. We wish to have a more thorough investigation on the differences between direct physical interactions and indirect functional associations,

however, due to the limitations of this study, we can only present the characteristics of the datasets, but can't achieve more precise quantitative analysis in this thesis. This Chapter thus provides an important guidance to the researchers who might base their works on the two *M. tuberculosis* H37Rv PPI datasets. The details of our analyses are presented in the sections below.

2.2 Method

2.2.1 Preparing STRING PPI datasets for analyses

STRING database uses a combination of prediction approaches and an integration of other information (neighborhood, transferred neighborhood, gene fusion, text mining, databases, homology transfer, cooccurrence, experiments and so on). STRING PPIs come from a mix of experimental data; PPIs copied from public databases (e.g. KEGG and BioGRID) and predicted PPIs. So we derive from STRING a subset of predicted PPIs and name this unbiased STRING subset "predicted functional associations dataset". This dataset is derived only from the following prediction approaches: neighborhood, transferred neighborhood, gene fusion, cooccurrence, transferred co-expression, text mining, and transferred text mining.

2.2.2 The agreement between a benchmark PPI dataset and a testing PPI dataset

We use Jaccard coefficient, recall, and precision to measure the agreement between a benchmark PPI dataset and a testing PPI dataset. Jaccard coefficient is defined as the size of the intersection of the two datasets divided by the size of the union of the two datasets. Recall is the proportion of benchmark PPIs that are in the testing dataset. Precision is the proportion of testing PPIs that are in the benchmark dataset. Thus,

1. Jaccard coefficient = TP / (TP+FP +FN);

2. Precision = TP / (TP+FP);

3. Recall = TP / (TP + FN).

Here, TP (true positives) represents the number of PPIs in the testing dataset that overlap with the benchmark dataset; FN (false negatives) represents the number of PPIs in the benchmark dataset that are not in the testing dataset; TN (true negatives) represents the number of all possible PPIs that appear in neither the testing dataset nor the benchmark dataset; and FP (false positives) represents the number of PPIs in testing dataset but are not in the benchmark dataset. The Jaccard coefficient, recall, and precision of the benchmark and testing datasets considered in this Chapter are given in Figures 2.1, 2.3, 2.5 and 2.6.

2.2.3 STRING score distribution of "Overlap PPI Number ratio"

In order to find which STRING score region has a higher percentage of overlapping PPIs with the B2H PPI dataset, STRING score distribution of "overlap PPI number ratio" between the STRING predicted functional associations dataset and the *M. tuberculosis* H37Rv B2H PPI dataset were calculated and plotted in Figure 2.2. At each score interval of 10, the "overlap PPI number ratio" is defined as the number of overlapping PPIs divided by the total number of PPIs in that interval. For example, if there are 300 PPIs from the STRING predicted functional associations dataset are in score range 150-160, and among these 300 PPIs there are 30 PPIs overlapping with the B2H PPI dataset, then in this score range 150-160 the "overlap PPI number ratio" is 30/300 = 0.1. We calculate all the "overlap PPI number ratio" in each interval, STRING score ranging from 150 to 1000, and the distribution of the ratios are plotted in Figure 2.2.

2.2.4 GO term annotation, informative GO term identification and PPI datasets assessments

M. tuberculosis H37Rv proteins are annotated with GO terms using InterProScan (Quevillon et al., 2005). GO terms are organized into three separate hierarchical ontologies—viz., cellular component terms (CC), molecular function terms (MF), and biological process terms (BP). A protein that is annotated by a particular GO term is considered to be annotated by all ancestor terms (in the corresponding hierarchical ontology) of that GO term—that is, the so-called "through-path" rule is applied. As top-level GO terms tend to be annotated to many proteins and leaf-level GO terms to very few proteins, in order to avoid bias in our analysis, we keep only "informative" GO terms for analysis.

A pair of proteins comes into contact with each other and interacts to perform a function. If the GO term annotations of the proteins in an organism are complete, we can expect such a pair of interacting proteins to have at least one informative GO term annotation in common. Therefore, a predicted or reported PPI is more likely to be a false positive when the two proteins in the PPI do not have an informative GO term annotated to them in common. We can therefore gauge the quality of a PPI dataset by calculating the percentage of PPIs in the dataset (where both proteins in the PPI have informative GO term annotations) that has "coherent" informative GO term annotations. A PPI is said to have coherent informative GO term annotation if the two proteins in the PPI have an informative GO term annotation in common.

However, the percentage of PPIs in a dataset that have coherent informative GO terms can be affected by the number of informative GO terms and by biases in the distribution of proteins these informative GO terms are annotated to. An informative GO term is defined as a GO term that has at least 30 proteins assigned to it or its descendants and none of its child terms have 30 or more proteins assigned to it. For example, if only one informative GO term was available in the organism, then 100%

of the annotated PPIs would be coherent. Thus, to better assess the quality of a PPI dataset by coherence of informative GO term annotations, we need to compare the percentage of coherently annotated PPIs in the dataset to appropriately generate random PPI datasets. In particular, a high ratio (named "Info GO ratio") of the percentage of coherently annotated PPIs in the PPI dataset compared to that of the random PPI dataset suggests that PPI dataset is likely to be of high quality. We generate random PPI network using the Random Network Plug-in in Cytoscape (Shannon et al., 2003). The percentage of PPIs that have coherent informative GO term annotations in the PPI datasets considered in this Chapter is given in Figures 2.3.

2.3 Result

This section can be divided into four parts: (i) Discover the low similarity between the two main H37Rv PPI datasets. (ii) Evaluate the quality of the two H37Rv PPI datasets in the same organism. (iii) Assess the quality of the H37Rv B2H PPI dataset across organisms. (iv) Analyse characteristics of the STRING PPIs in *M. tuberculosis* and *S. cerevisiae*.

2.3.1 Lack of agreement between the two *M. tuberculosis* H37Rv PPI datasets

The H37Rv B2H PPI dataset is used as benchmark and different subsets of H37Rv STRING PPIs are tested against it. We consider all H37Rv STRING PPIs as well as STRING H37Rv PPIs based on specific methods (gene neighbourhood, gene fusion, etc.) used for predicting them. For each subset, in Figure 2.1, we show the Jaccard coefficient, precision, and recall of each predicted subset at STRING score ≥ 770.

According to Figure 2.2, STRING score threshold at around 770 generally maximizes the overlap between two-hybrid PPIs and STRING predicted PPIs in *M. tuberculosis* H37Rv. It is clear from Figure 1 that the STRING PPIs predicted by various

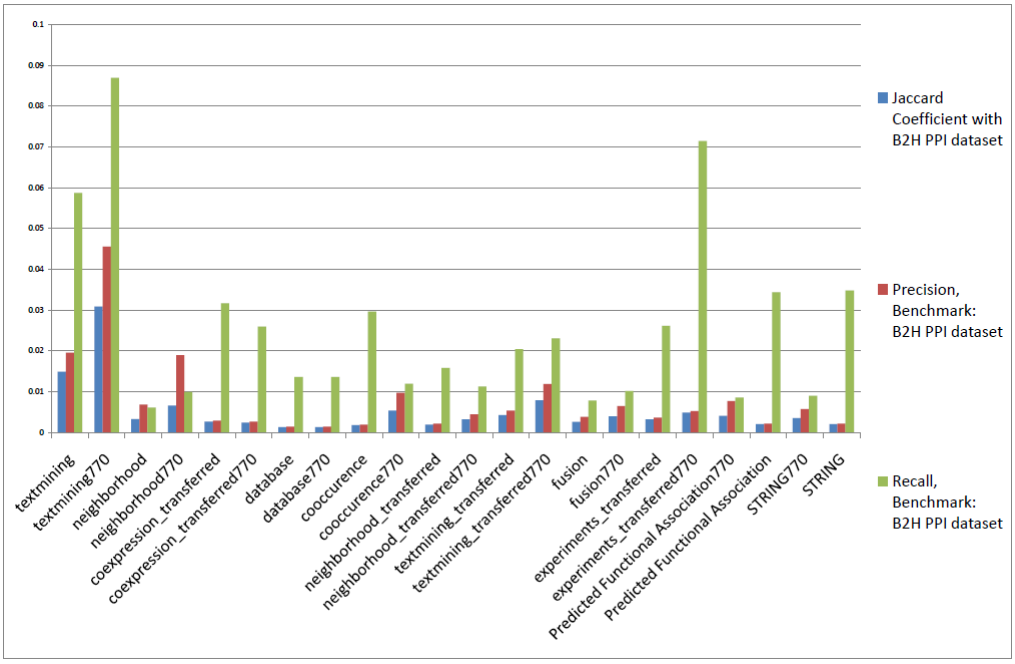


Figure 2.1: Agreement between H37Rv PPIs in STRING and the B2H PPI datasets. The Jaccard coefficient, precision and recall between H37Rv PPI datasets in STRING database predicted by different methods and the H37Rv B2H PPI dataset (benchmark).

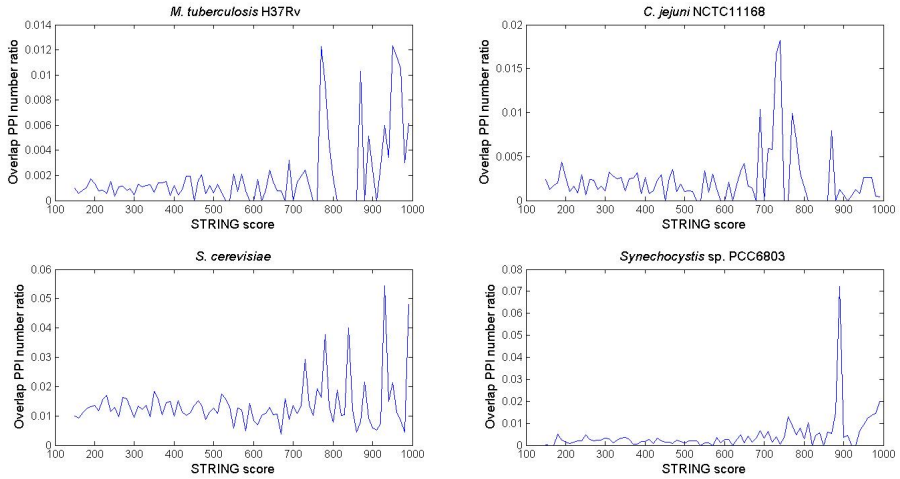


Figure 2.2: Overlap PPI number ratios at various STRING score thresholds. The overlap PPI number ratios at various STRING score thresholds between (i) the H37Rv B2H PPI dataset and the H37Rv STRING predicted functional associations dataset, (ii) the *S. cerevisiae* Y2H PPI dataset and the *S. cerevisiae* STRING predicted functional associations dataset, (iii) the *C. jejuni* NCTC11168 Y2H PPI dataset and the *C. jejuni* NCTC11168 STRING predicted functional associations dataset, and (iv) the *Synechocystis* sp. PCC6803 Y2H PPI dataset and *Synechocystis* sp. PCC6803 STRING predicted functional associations dataset.

methods all have extremely low precision, recall, and overlap with the H37Rv B2H PPI dataset. Below are some representative statistics:

1. Overlapping PPIs between the two datasets: 276
2. STRING PPIs Precision: 0.00215 Recall: 0.03503
3. STRING PPIs (at score $\geq$ 770) Precision: 0.00574 Recall: 0.00896

The extremely low agreement between the H37Rv PPIs in the STRING and B2H PPI datasets is rather unexpected. We hypothesize that it may be a result of one or more of the following situations. First, it may be that the H37Rv B2H PPI dataset contains an unusually high level of noise. Second, it may be that the H37Rv STRING PPI dataset and subsets thereof contain an unusually high level of noise. Third, it may be that the predicted PPIs in STRING are not direct physical interactions; rather, they may primarily be other types of functional associations such as protein pairs in the same protein complexes and enzyme pairs catalyzing successive reaction steps.

2.3.2 Overlap PPI number ratios at various STRING score thresholds

The results above reveal the surprisingly low coverage between the two H37Rv PPI datasets. However, as shown in Figure 2.2, at STRING score $\geq$ 770, there is a higher level of overlap between the two datasets. This increase in overlap between two-hybrid PPI dataset and STRING predicted functional associations dataset at high scores is also observed in *C. jejuni* (Parrish et al., 2007), *Synechocystis* (Sato et al., 2007) and *S. cerevisiae* (Yu et al., 2008). This suggests that STRING PPIs with high score potentially has higher quality than STRING PPIs with a lower score. Nevertheless, the overlap between these two-hybrid PPI datasets and their respective STRING predicted functional association datasets is no more than 8% at any score interval. Thus, even at a high STRING score threshold, there is no clear agreement between two-hybrid PPIs and STRING predicted functional association datasets. Assuming that not all of

these two-hybrid PPI datasets are of low quality, this lack of clear agreement strongly suggests that STRING predicted PPIs are unlikely to correspond mainly to direct physical interactions.

2.3.3 Assessment of PPI datasets using informative GO terms

Two interacting proteins are more likely to be localized in the same cellular component and/or having a common function role than not (Chua et al., 2006). So we calculate the percentage of PPIs in a PPI dataset having coherent informative GO terms—i.e., the rate of interacting protein pairs with common function roles (measured based on informative GO terms in MF and BP categories) and cellular localization (measured based on informative GO terms in the CC category) in the PPI dataset—to evaluate the quality of the PPI dataset.

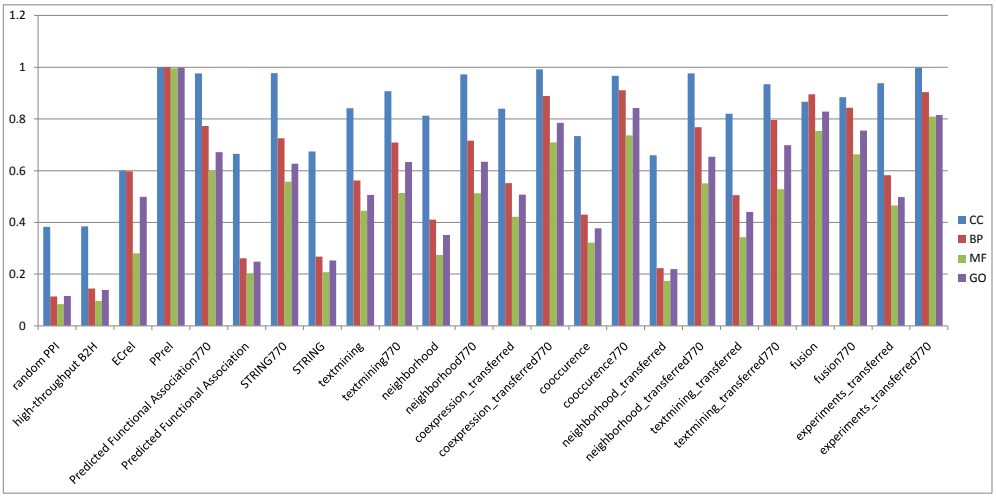


Figure 2.3: Percentage of PPIs in various *M. tuberculosis* PPI datasets that have coherent informative GO term annotations. Percentage of PPIs in various *M. tuberculosis* PPI datasets that have coherent informative GO term annotations.

The percentage of PPIs having coherent informative GO terms is computed for each of the datasets in Figure 2.3. This way, exactly one GO term is considered in any

through path. Moreover, each GO term considered is at the finest resolution possible while being annotated to a sufficiently large number of proteins (≥ 30) for a valid analysis. The datasets include subsets of STRING derived from specific source channels in STRING. Note that some source channels may introduce GO-related information into STRING. In particular, the "database" and "database transfer" channels may collect PPIs derived from Protein Complexes in the Gene Ontology (GO) database. Thus, to avoid circularity in our results here and elsewhere, we mainly use statistics from an unbiased subset "predicted functional associations dataset" of STRING obtained by excluding the PPIs from source channels that may introduce confounding factors. The "predicted functional associations dataset" consists of STRING PPIs that are generated only from the following prediction approaches: gene neighborhood, transferred neighborhood, gene fusion, and co-occurrence, transferred co-expression, text mining, and transferred text mining. Among the three categories of GO terms, the datasets generally show a high percentage of coherence with respect to informative CC GO terms. However, this observation should be dismissed because there are only three distinct informative CC GO terms, which is an order of magnitude less than informative MF and BP GO terms. The random PPI dataset has the lowest percentage of PPIs with coherent informative GO terms in all the tested PPI datasets, which makes sense. But the H37Rv B2H PPI dataset has the second lowest percentage of PPIs with coherent informative GO terms and is very close to the random PPI dataset. This indicates that the *M. tuberculosis* H37Rv B2H PPI dataset has the lowest quality among all the PPI datasets been evaluated. The H37Rv predicted functional association dataset (without thresholding at score ≥ 770) also has a low percentage of PPIs with coherent informative GO terms and is thus of low quality. However, most PPI datasets from STRING (with score ≥ 770) show a much higher percentage of PPIs having coherent informative GO terms than the H37Rv B2H PPI dataset, suggesting that a higher percentage of PPIs in these datasets may have better reliability than those of the B2H PPI dataset

and of the STRING PPI dataset as a whole.

2.3.4 Analysis of PPI datasets using gene expression profile correlation

Two proteins that interact are more likely to be correlated in the expression of their underlying genes than not (Grigoriev, 2001). In fact, co-expression is one of the prediction methods in STRING (Szklarczyk et al., 2011). However, there is no STRING PPI predicted from co-expression in *M. tuberculosis* H37Rv. Given that *M. tuberculosis* H37Rv gene expression data is readily available in public repositories, this lack of H37Rv PPIs predicted using this information is an unexpected limitation of STRING. At the same time, this absence makes using correlation of gene expression profiles for assessing the quality of the H37Rv B2H and STRING PPI datasets unbiased. The results in Figure 2.4 clearly show that the H37Rv STRING PPI dataset (at score $\geq$ 770) has a much larger proportion of PPIs that exhibit correlation in the expression profiles of their underlying genes than the H37Rv B2H PPI dataset and the whole H37Rv STRING PPI dataset. In fact, a mere 223 PPIs in the H37Rv B2H PPI dataset have significant correlated gene expression profiles (Pearsons correlation coefficient >0.4). These 223 PPIs are likely to be more reliable than most of the other PPIs in the H37Rv B2H dataset.

2.3.5 Analysis of the characteristics of *M. tuberculosis* H37Rv PPIs using pathway gene relationships

From the results presented earlier, it seems that many H37Rv STRING PPIs may not be direct physical interactions. In order to understand what these PPIs may better correspond to, we collect pair-wise protein/gene relationships from several major pathway databases, and compare them with the various PPI datasets considered earlier in this paper. The ECrel dataset comprises enzyme pairs that catalyze successive reaction

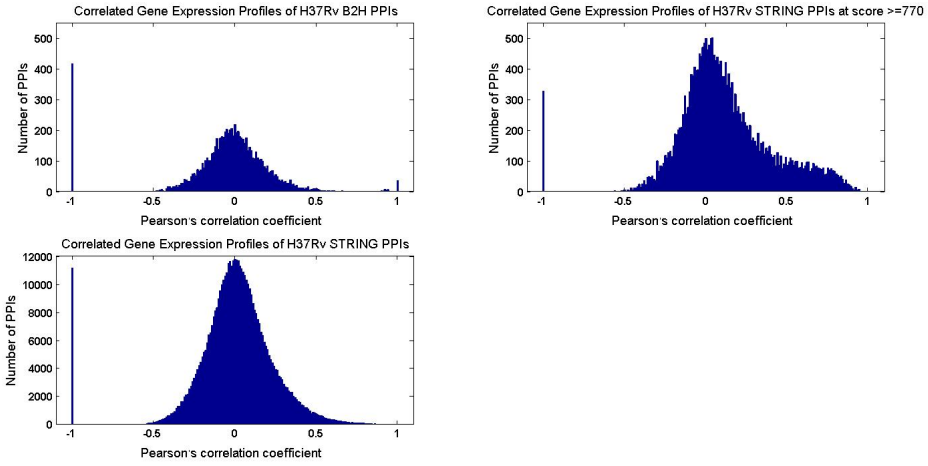


Figure 2.4: PPI datasets assessment by gene expression profile correlation. The distribution of Pearsons correlation coefficient of the expression profiles of underlying genes of different PPI datasets are given in this figure (x axis is the Pearsons correlation coefficient, y axis is the number of PPIs). The bar at -1 in the charts here corresponds to PPIs where we do not have the expression profiles of their underlying genes.

steps in enzymatic pathways. The PPrel dataset comprises more direct protein-protein interactions but it also contains protein pairs in the same complexes. Thus, a PPI dataset containing more indirect protein relationships should show high similarity to the ECrel dataset.

However, this task is hampered by the sparse information stored in all the current main pathway databases, like KEGG, WikiPathways and BioCyc. Therefore an integration of pathway information from the three main databases is needed to maximize the effectiveness of pathway information for this comparative analysis of PPI datasets. We calculate the Jaccard coefficient, precision, and recall of each of the *M. tuberculosis* H37Rv PPI datasets discussed earlier using ECrel (Figure 2.5) from *M. tuberculosis* H37Rv integrated pathway gene relationships as the benchmarks.

Results from above experiments show that the *M. tuberculosis* H37Rv B2H PPI dataset shows very low similarity with ECrel dataset, while most of STRING PPI datasets show good similarity. This provides another explanation for the low similarity

between the H37Rv B2H and STRING PPI datasets. Namely, the former dataset contains direct physical interactions, as it is to be expected of B2H assays; while the latter STRING datasets also include substantial amounts of PPIs that are indirect protein relationships.

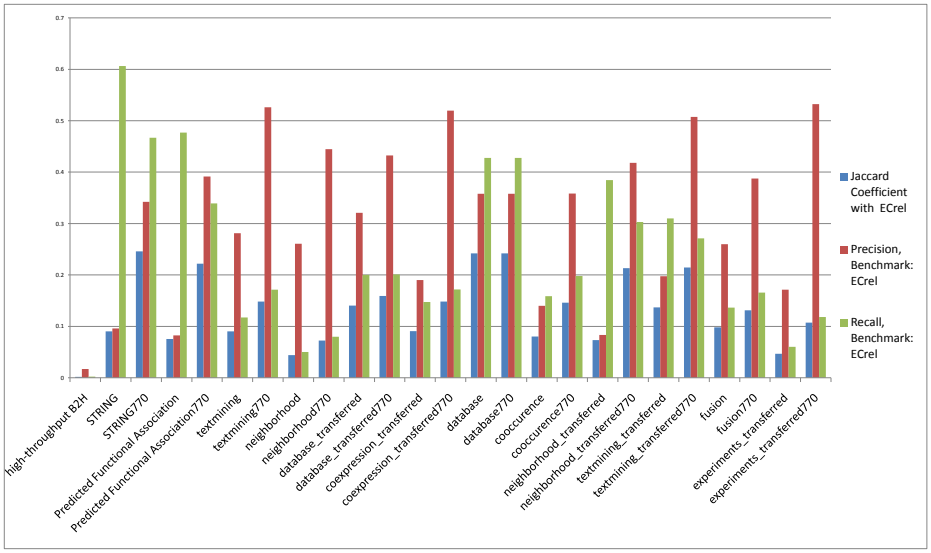


Figure 2.5: Comparative analysis of PPI datasets using integrated pathway gene relationships (ECrel). *M. tuberculosis* H37Rv PPI datasets similarity to integrated pathway gene relationships (ECrel dataset as benchmark).

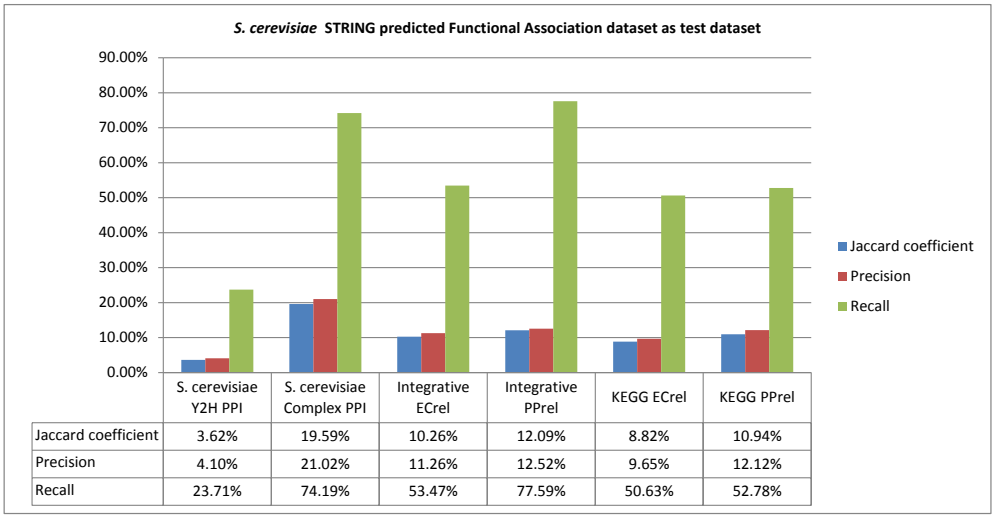
2.3.6 STRING PPI dataset analysis in *S. cerevisiae*

The comparative analysis of the various H37Rv PPI datasets using integrated pathway gene relationships reveals that the H37Rv STRING PPI dataset may contain a lot of indirect protein relationships. The STRING database has proclaimed itself as a database consisting of "known and predicted protein-protein interactions" (Szklarczyk et al., 2011). In practice, both physical interactions and functional associations, and both predicted and experimental "PPIs" are included in this database. Therefore, it is important to clearly demonstrate which kind of PPIs are contained in STRING. We return to the most comprehensively investigated model organism—*S. cerevisiae* —to

more precisely analyze the characteristics of PPIs in STRING. As a unified database, the PPIs prediction approaches in the STRING database are consistently used on all the 630 organisms included in it. Thus the phenomena discovered in *M. tuberculosis* H37Rv should also exist in other organisms like *S. cerevisiae*, and vice versa. Moreover, we have much more information in *S. cerevisiae* that can be used for conducting a much more precise analysis. If the situation observed earlier that the *M. tuberculosis* H37Rv STRING PPI dataset contains a lot of indirect PPIs is also observe in *S. cerevisiae*, then it will be a sound confirmation of our earlier conclusion.

We similarly obtain the integrated pathway gene relationships (mainly ECrel and PPrel) for *S. cerevisiae* and also separate datasets prepared only from KEGG for more precise reference. We further collect all protein pairs (named the "*S. cerevisiae* Complex PPI dataset") that appear in the same protein complexes using the protein complexes dataset from Wodak Lab (Pu et al., 2009). It is obvious that the "*S. cerevisiae* Complex PPI dataset" may contain a lot of indirect PPIs, like relationships between two non-directly-binding proteins in protein complexes. A representative *S. cerevisiae* two-hybrid PPI dataset (Yu et al., 2008) is also included in this comparative analysis. To avoid a biased comparison, as the full STRING PPI dataset may include many PPIs from the datasets above, we use the *S. cerevisiae* predicted functional associations dataset from STRING database as the testing dataset in this analysis. The overlapping number of PPIs, Jaccard coefficient, precision and recall are calculated, and the results are given in Figure 2.6. From the results, the *S. cerevisiae* two-hybrid PPI dataset has the lowest similarity to the *S. cerevisiae* STRING predicted functional associations dataset, whereas the complex PPI dataset and ECrel datasets (both from KEGG and from integrated pathway gene relationships) reveal good similarity to the *S. cerevisiae* STRING predicted functional associations dataset. This result is in accordance with the results on *M. tuberculosis* H37Rv, clearly demonstrating that the STRING database PPIs include a substantial amount of PPIs that are indirect protein relationships,

including protein pairs in the same protein complexes and protein pairs catalyzing successive enzymatic reaction steps.



	S. cerevisiae Y2H PPI	S. cerevisiae Complex PPI	Integrative ECrel	Integrative PPrel	KEGG ECrel	KEGG PPrel
Jaccard coefficient	3.62%	19.59%	10.26%	12.09%	8.82%	10.94%
Precision	4.10%	21.02%	11.26%	12.52%	9.65%	12.12%
Recall	23.71%	74.19%	53.47%	77.59%	50.63%	52.78%

Figure 2.6: Comparative analysis of different *S. cerevisiae* protein relationships datasets with *S. cerevisiae* STRING functional associations dataset. Comparison of the similarity between different protein relationships datasets with *S. cerevisiae* predicted functional associations from STRING database.

2.4 Discussion

2.4.1 Reliable *M. tuberculosis* H37Rv B2H PPI datasets

We have shown that the *M. tuberculosis* H37Rv B2H PPI dataset has low quality. In the process, we find four subsets of the B2H PPI dataset that may be more reliable than the rest of this dataset. The first subset consists of PPIs where both interaction partners have coherent informative GO terms—viz., B2H PPIs sharing functional homogeneity or localization coherence. This subset contains 115 PPIs and is named "B2HsameGO dataset". The second contains overlapping PPIs between the H37Rv B2H and STRING PPI datasets, which can be considered as B2H PPIs supported by STRING prediction approaches. This subset consists of 276 PPIs and is named "B2HSTRING dataset".

The third subset contains those H37Rv B2H PPIs that have been verified by different experiments. This subset has 147 PPIs and is named "B2HdiffExp dataset". The fourth subset contains PPIs where both interaction partners have significant correlated gene expression profile (Pearsons correlation coefficient >0.4). This subset has 223 PPIs and is named "B2Hco-express dataset". The overlap between any pair of these four subsets is small because the PPIs in these four subsets involve very different proteins.

There are several inherent limitations of PPI data generated by two-hybrid approaches (both B2H and Y2H), including PPIs that are detected between over-expressed proteins, between fusion proteins, and in a different host (Yeast or *E. coli*). Given the data available in this study, we are not able to clearly identify which erroneous B2H PPIs are caused by which inherent limitations of the two-hybrid system. We leave this interesting and difficult challenge to a future project.

2.4.2 Differences between functional associations and physical interactions

Physical interactions correspond to direct protein relationships such as two proteins binding to each other. Functional associations can be both direct and indirect protein relationships; for example, two enzymes catalyzing successive reaction steps can be regarded as functional associations. This partially explains the differences between the H37Rv B2H and STRING PPI datasets, as we have demonstrated earlier.

Based on the approach used in generating the PPIs, each of the two major categories can be further divided into two parts, experimental physical interactions (e.g., PPIs from Y2H or co-purification); predicted physical interactions (e.g., interologs predicted from co-purification PPIs); experimental functional associations (e.g., PPIs from synthetic lethality or dosage growth defect); predicted functional associations (e.g., PPIs from neighbourhood or co-occurrence). Differences among PPI datasets from the four categories are inevitable, and they all have some portion of real PPIs and some noise.

However, a high noise level often overwhelms the agreement between the datasets from these four categories. Real PPIs are both functional associations and physical interactions (intersect dataset), because if two proteins truly interact with each other in normal environments, the two proteins must have functional relationships. The four subsets of reliable H37Rv B2H PPIs identified by us only contain a small number of PPIs and are not enough to illuminate the whole direct physical interactome in *M. tuberculosis* H37Rv.

2.5 Conclusions

In this Chapter, we have observed the strikingly low agreement between *M. tuberculosis* H37Rv B2H and STRING PPI datasets. We have demonstrated the two main causes of this low level of agreement. The first reason is the low quality of the B2H PPI dataset, which seems to contain a significant amount of false positives as well as false negatives. The same is true of the H37Rv STRING PPI dataset as a whole, though a subset comprising PPIs with score ≥ 700 seems more reliable. The second reason is that the STRING PPI dataset contains a substantial amount of predicted PPIs that are not direct interactions, which seem more likely to correspond to protein pairs that are in the same protein complexes or protein pairs that are catalyzing adjacent reaction steps in enzymatic pathways. Because of the low quality of the H37Rv B2H PPI dataset, it should not be used as a gold standard to evaluate the quality of other *M. tuberculosis* PPI datasets, predicted or otherwise. Researchers who need to use this dataset should do so with great caution. Yet, as the only available large-scale physical interaction dataset of *M. tuberculosis* H37Rv at the moment, even though it suffers from high noise and low quality, the direct protein physical interaction information in this dataset should not be ignored. We have identified four subsets of this B2H PPI dataset that are more reliable, which can be combined into a single dataset, which can serve as a suitable reference H37Rv physical interaction dataset for many applications.

STRING score is useful for indicating which STRING PPIs have higher quality. We suggest a STRING score threshold set around 770. Therefore, this dataset can be a good source as a functional associations reference in *M. tuberculosis* H37Rv.

Chapter 3

IntPath—Integration and Database

Pathway data are important for understanding the relationship between genes, proteins and many other molecules in living organisms. Many well-established databases—e.g., KEGG , WikiPathways, and BioCyc—are dedicated to collecting pathway data for public access. However, the effectiveness of these databases is hindered by issues such as incompatible data formats, inconsistent molecular representations, inconsistent molecular relationship representations, inconsistent referrals to pathway names, and incomprehensive data from different databases.

We have overcome in IntPath the issues of compatibility, consistency, and comprehensiveness that often hamper effective use of pathway databases. We have included four organisms in the current release of IntPath. Our methodology and programs described in this Chapter can be easily applied to other organisms; and we will include more model organisms and important pathogens in future releases of IntPath. IntPath maintains regular updates and is freely available at `http://compbio.ddns.comp.nus.edu.sg:8080/IntPath`.

3.1 Background

The proliferation of pathway databases—e.g., KEGG (Ogata et al., 1999), WikiPathways (Pico et al., 2008; Kelder et al., 2009), BioCyc (Karp et al., 2005; Karp, 2001), and MouseCyc (Evsikov et al., 2009)—are useful for understanding the relationship between genes, proteins and other molecules in living organisms. However, the effectiveness of these databases is hindered by issues such as incompatible data formats, inconsistent molecular representations, inconsistent molecular relationship representations, inconsistent referrals to pathway names, and incomprehensive data from different databases. These difficulties call for an effective integration of these databases.

There are many approaches to integrate pathways. For example, Pathway Commons and PathCase (Elliott et al., 2008) can be considered as taking the "aggregator" approach. In this approach, a common access method and data format are adopted or developed for a set of pathways imported from a collection of source databases. The aggregator approach does not perform any unification of the underlying pathways—viz., if n source databases each contains information on a particular pathway, that pathway is presented by the aggregator as n separate pathways. On the other hand, GenMapp (Salomonis et al., 2007), Cytoscape (Shannon et al., 2003), and PathVisio (van Iersel et al., 2008) can be considered as taking the "converter" approach. Basically, these tools support the import and export of biological pathways in a variety of formats, even though these tools are designed mainly for exploring, visualizing, and editing biological pathways. Lastly, PathwayAPI (Soh et al., 2010) can be considered as taking the "full unification" approach. In this approach, pathways in different source databases that are meant to represent the same pathway are merged and molecular objects mentioned in the different source pathways that are meant to represent the same objects are matched. This approach is technically more difficult than other approaches; but it has the advantage of presenting a more coherent and comprehensive view of the pathways.

Very recently, Stobbe et al. (2011) compared the genes, EC numbers and reactions of five frequently used human metabolic pathway databases. They found that the overlap between these databases is surprisingly low. More importantly, their results show that each of the five networks compared provides a valuable piece of the puzzle of the comprehensive reconstruction of the human metabolic network. This discovery is a strong motivation for the "full unification" approach mentioned above. Stobbe et al. further suggested that, for an effective integration, one needs to standardize the metabolite names and identifiers and to resolve the conceptual differences between the databases.

Besides the databases that focus specifically on pathway data integration, some protein functional interaction databases have also extended their collection to pathway data. For example, ConsensusPathDB (Kamburov et al., 2011) integrates different types of functional interactions from heterogeneous interaction data resources and pathway databases for three organisms(human, yeast and mouse).The distinct difference in their primary focus results in an obvious difference between ConsensusPathDB and IntPath. ConsensusPathDB collects pathway data from many databases but does not appear to produce integrated pathways—even when the same pathway is present in different sources, they are still listed individually without merging. How to merge the different instances of the same pathways among and within the source pathway databases is the major concern of IntPath. Unlike ConsensusPathDB, IntPath mainly focuses on the integration of pathway-gene and pathway-gene pair relationships , with the aim of solving the problem of inconsistencies and incomprehensiveness among different pathway databases.The definition of "gene pair" in this Chapter is the gene-gene relationship in pathways, the relationship type of the two components in each gene pair is described in table 3.1.

In this Chapter, we take this full unification approach in building **IntPath**, the **Int**egrated **Path**way gene relationship database for model organisms and important

Unified Genes relationships	Explanation
ECrel	Enzyme-enzyme relation, indicating two enzymes catalyzing successive reaction steps.
PPrel	Protein-protein interaction, such as binding and modification, or proteins have control over the same process.
GErel	Gene expression interaction, indicating relation of transcription factor and target gene product.
GPrel	Proteins belong to the same molecular complex, not necessarily interacting directly.

Table 3.1: Four types of IntPath unified gene relationships. Explanations of the types of relationships in IntPath are given below.

pathogens. This approach was also taken earlier by Soh et al. (2010) when they integrated general human pathways into PathwayAPI. IntPath differs from PathwayAPI in several aspects. In terms of content, a different set of databases and multiple organisms are considered in IntPath. In terms of data extraction, IntPath extracts all pathway data directly from the xml files of each source database and the whole process is highly automated. Therefore, IntPath provides integrated and unified pathway information on a much larger set of organisms and it can be extended to include many other organisms in a short time. In contrast, PathwayAPI integrated only human pathways. Also, for all the organisms included in IntPath, a regular update of each organism can be maintained. In terms of pathway data integration, IntPath not only looks for related pathways between databases but also within each source database; this integration approach provides more unified, meaningful and comprehensive integrated pathway-gene and pathway-gene pair relationships information. In contrast, PathwayAPI only looks for related pathways between databases but not within the same source database. Moreover, IntPath also provides more features and tools. It not only supports web service but also a full-featured web interface. More analysis tools based on pathway data have been provided—like “Analyze Distance” and “Identify Pathways”—and more analysis functions and tools will continue to be added on IntPath in future releases.

The incompatible data formats of different databases seriously inhibit effective and compatible information retrieval. In KEGG, pathways are represented in KGML format and SOAP (returned when using API calls). In WikiPathways, pathways are represented in GPML format; recently it begins to support web service (Kelder et al., 2009), allowing users to access the data through API calls; and the BioPAX format is also supported. In BioCyc and MouseCyc, the pathway data are primarily represented in the BioPAX format. In IntPath, we overcome this limitation by extracting the pathway gene relationships from these different databases and convert these various complicated XML-based formats into simple tab-delimited text files.

Inconsistent molecular representations significantly lower the effectiveness of pathway information retrieval. Different databases maintain different naming conventions on the nodes of their pathways. In KEGG, the names of the nodes (genes and proteins) in the pathways can be KEGG Entry name, KEGG ORTHOLOG (KO) ID, etc. The graphic names on KEGG pathway map can be Gene Symbol (or synonym), Enzyme Commission (EC) number, etc. In WikiPathways, the nodes' "TextLabel" are given gene symbol (or synonym), gene name, protein name, EC number, etc. In most cases the nodes can also be given Entrez ID, NCBI Accession, Ensembl Gene ID, Ensembl protein ID, UniProt Assession, etc. And some times nodes in WikiPathways are only given "TextLabel" without any database reference ID. MouseCyc (Evsikov et al., 2009) mainly uses MGI ID and also includes the corresponding gene symbol, UniProt accession, etc. BioCyc (MTBRvCyc, YeastCyc and HumanCyc) Accession Number is mainly used to represent nodes in the pathways while corresponding gene symbol, gene name (protein name), Entrez ID, and UniProt accession number are sometimes included.

Inconsistent molecular relationship representations may also cause confusion when referencing pathway information from different repositories. In KGML (KEGG), the relationships between molecules are represented as PPrel, ECrel, PCrel, GErel, etc.

In GPML (WikiPathways), the relationships can be inhibition, activation, protein complexes, enzyme complexes, acetylation, phosphorylation, etc. In BioPAX (BioCyc and MouseCyc), when transformed into the SIF format, the relationships can be SEQUENTIAL_CATALYSIS, CO_CONTROL, INTERACTS_WITH, etc. These inconsistencies cause troubles for researchers wishing to refer to pathway information in a large-scale manner across different databases. Therefore, some normalization technique is needed to convert the nodes and edges from different pathways in different repositories into a common representation. In IntPath, we overcome the above two limitations by normalizing the pathway gene representations and gene relationship representations from different databases into unified IntPath gene and relationship representation. The unified IntPath gene ID for *Homo Sapiens* is HGNC Symbol, *Mus musculus* is MGI Symbol, *Saccharomyces cerevisiae* is Systematic name, and *Mycobacterium tuberculosis* H37Rv is TuberList Rv number. The unified IntPath gene relationship representations are listed in Table 3.1.

Inconsistent referrals to pathway names are another source of confusion that substantially reduces the effectiveness of retrieving information on the same pathway from different databases. For instance, KEGG may refer to a pathway as "Glycolysis / Gluconeogenesis", and WikiPathways may name it as "Glycolysis and Gluconeogenesis". For another example, WikiPathways contains a pathway with the name "Cholesterol Biosynthesis", while BioCyc has many corresponding pathways such as "cholesterol biosynthesis III (via desmosterol)", "cholesterol biosynthesis II (via 24, 25-dihydrolanosterol)", "cholesterol biosynthesis I", and "superpathway of cholesterol biosynthesis". Therefore, a unified pathway naming system may reduce the confusion when referring to the same or similar pathway information from different databases.

Furthermore, the comprehensiveness of data from different databases is another limitation of these pathway databases. By the term "incomprehensiveness", we mean that each single biological database is not a comprehensive representation of biological

knowledge that is considered by experts to be accurate (Soh et al., 2010). We reveal the incomprehensiveness of current databases via analysis on the agreement of the common pathway between these different databases. In IntPath, these inconsistencies and incomprehensiveness issues are solved by the integration approach.

3.2 Data

We choose several representative data sources—KEGG (Ogata et al., 1999), WikiPathways (Pico et al., 2008; Kelder et al., 2009), BioCyc (Karp et al., 2005; Karp, 2001), and MouseCyc (Evsikov et al., 2009)—for our analysis and integration. These data sources are selected because they are representatives of very different kinds of curation efforts. Currently, the following organisms are included in our IntPath database (version 2.0): *Homo sapiens*, *Mus musculus*, *Saccharomyces cerevisiae* and *Mycobacterium tuberculosis* H37Rv. For each organism included in IntPath, the pathway data are collected from three representative databases: 1. KEGG; 2. WikiPathways; 3. One of the following four databases—YeastCyc, HumanCyc (Romero et al., 2005), and MTBRvCyc from the BioCyc collection (Karp et al., 2005; Karp, 2001); and MouseCyc (Evsikov et al., 2009). The four Pathway/Genome Databases (PGDBs)—MouseCyc, YeastCyc, HumanCyc, MTBRvCyc—are generated and recorded in a very similar way, but the PGDBs of different organisms are maintained and curated by different groups.

MouseCyc is curated by the Jackson Laboratory; it is a new, manually curated database of both known and predicted metabolic pathways for the laboratory mouse (Evsikov et al., 2009). YeastCyc is a Tier-1 PGDB from the BioCyc collection (Karp et al., 2005; Karp, 2001); it is curated by SGD Curators in Stanford University. PGDBs in Tier 1 have received more than one year of literature-based curation by scientists. MTBRvCyc and HumanCyc (Romero et al., 2005) are Tier-2 PGDBs from the BioCyc collection; they are generated by the PathoLogic program and received moderate curation (mostly have undergone 1-4 months of curation). WikiPathways is maintained by

a community of users via a wiki-style platform (Pico et al., 2008; Kelder et al., 2009). KEGG database is curated independently by a single lab from published literature (Soh et al., 2010).

3.3 Methods

3.3.1 Extraction and normalization of pathway-gene and pathway-gene pair relationships

The first step of extracting information from pathway databases is downloading the XML files. To automatically download the hundreds of KGML files of each organism on the KEGG ftp site, we use a simple spider program written in Perl. For BioCyc and MouseCyc, the BioPAX files—and for WikiPathways, the GPML files—are compressed into a single package which can easily be downloaded manually.

Extracting the pathway-gene and pathway-gene pair relationships from KGML is accomplished using an in-house Java program, which extensively uses regular expressions to retrieve specific information from the KGML files. A KGML file consists of entries like "</entry>", "</relation>" and "</reaction>"; in each entry there is either entry information of the nodes (genes, enzymes, compounds , ortholog groups and so on), groups (complexes of gene products like protein complexes and so on) or relationships (relationship between the nodes in the pathway map). Using regular expressions we can specifically obtain the genes of each pathway and the relationships between each gene, and then link these genes according to the relationships. For genes belonging to complexes (groups), the binary gene pairs are generated based on the matrix model.

An alternative way of retrieving KEGG pathway genes and gene pairs is by calling the KEGG API, which enables users to easily use their programs to get access to the KEGG database. However, the API is not well updated (Soh et al., 2010). The KEGG

API does provide a function that can retrieve gene relationships from this database, though the results returned are KGML entry IDs, not exactly as we have wanted. Although calling the KEGG API would work in theory as described in (Soh et al., 2010), we turn to mining KGML directly and achieve the same good results.

Extracting pathway-gene and pathway-gene pair relationships from a GPML file is also accomplished using a strategy similar to mining KGML files. Mining a GPML file is much more difficult due to its wiki-style; and there are slight variations among individual GPML files even in the same organism, like some key tags may be in upper case in one file but lower case in another, random insertions of whitespace character, etc. Due to these variations, the regular expressions used for performing the extraction must be very robust. In GPML files, the information of genes and proteins are stored in a "</DataNode>" entry where, if the node is a gene or protein, the Type of the entry is set to "GeneProduct". The information of relationships (like activation and inhibition) are stored in a "</Line>" entry.

The linkage of "</DataNode>" entry and "</Line>" entry is mainly accomplished by their "graphID". A "</Line>" entry usually records which two "</DataNode>" entries it links to through the records of two corresponding "graphID" of the "</DataNode>" entries. Using this information we can retrieve the relationship of two genes linked by a "</Line>" entry. This relationship—like inhibition, activation, etc.—can be regarded as equivalent to the "PPrel" relationship in KGML.

If the genes belong to certain complexes (groups), their "GroupRef" ID are recorded in some "</DataNode>" entries; and genes with same "GroupRef" ID are in the same group (molecular complexes). All possible pair-wise relationships among members of a group are generated based on the matrix model. These relationships, derived from such a group, are mainly binary relationships among members in a protein complex or enzyme complex. They can thus be regarded as equivalent to the "group" relationship in KGML. This strategy works well for most GPML files; but, for some individual files,

there is simply no "graphID" in the file and, only positional information of each entry is recorded. This causes difficulty in retrieving the corresponding gene relationships. Attempts have been made to retrieve the pair-wise relationships based purely on the positional information; but these attempts also introduce a substantial amount of noise. Therefore we do not use this noisy information.

Retrieving gene relationships from BioPAX files is mainly by using the Paxtools Java programming library (Demir et al., 2010) in combination with our own simple Java program. By transforming a BioPAX (Level 2) file into the SIF format, we get both a Node file and an Edge file. Then a simple node mapping is made to retrieve gene relationships. These pair-wise relationships have no indication of the source pathway name. We need to map these relationships to their corresponding pathways. MouseCyc and BioCyc provide a file that clearly records all the genes in each pathway. Using this information, we are able to map the gene relationships to their corresponding pathways.

Converting the relationship of genes in complexes (groups) into binary gene pair relationships may not be the most ideal format for some users, who wish to refer to the original protein complexes information in pathways. For KEGG and WikiPathways we also maintain a "group-gene list" which specifically retains the original format of genes in the groups. The groups in this "group-gene list" are not integrated, as we have done for the pathway-gene and pathway-gene pair relationships, since maintaining this "group-gene list" is mainly to prevent information loss and to give users more precise original information that may not be easily reconstructed from the integrated pathway-gene pair relationships. We normalize the gene IDs in the "group-gene list" to IntPath unified gene IDs and store the list in a simple text file. Users can download this list along with other pathway-gene and pathway-gene pair relationships in the form of compressed text files from IntPath.

Normalization of gene names is done using gene ID mapping files downloaded from a variety of databases including NCBI(Maglott et al., 2005), KEGG (Ogata et al.,

H. sapiens	KEGG	WikiPathways	HumanCyc
Pathways	237	135	290
Genes	5,935	3,445	1,082
Gene Pairs	29,566	18,035	5,961
M. musculus	KEGG	WikiPathways	MouseCyc
Pathways	218	140	323
Genes	6,306	4,084	1,194
Gene Pairs	32,235	25,004	10,792
S. cerevisiae	KEGG	WikiPathways	YeastCyc
Pathways	98	125	184
Genes	1,735	863	542
Gene Pairs	2,922	57	1,440
M. tuberculosis H37Rv	KEGG	WikiPathways	MTBRvCyc
Pathways	110	8	234
Genes	1,078	152	493
Gene Pairs	3,775	62	2,764

Table 3.2: The number of pathways, genes and gene pairs from different databases after normalization. Summary of the number of pathways, genes, and gene pairs after normalization from different databases.

1999), UniProt (The UniProt Consortium, 2012), HGNC (Seal et al., 2011), MouseCyc (Evsikov et al., 2009), BioCyc (Karp et al., 2005; Karp, 2001), and BioMart (Haider et al., 2009). The gene relationships from different databases are mapped to the IntPath unified relationships listed in Table 3.1.

3.3.2 Evaluation of normalized pathway genes and gene pairs from different databases

After we have obtained the pathway-gene and pathway-gene pair relationships from different pathway databases, agreement among the databases can be analyzed. These agreement analyses are crucial for the downstream applications of IntPath. We examine three aspects of the agreement among the different pathway databases: (i) agreement of genes and gene pairs in different databases, (ii) agreement of the pathways in different databases, and (iii) agreement of genes and gene pairs of the same pathway in different databases.

After normalization the statistics about pathway number, gene number and gene

H. sapiens	KEGG vs WikiPathways	WikiPathways vs HumanCyc	HumanCyc vs KEGG
Overlap Genes	2,485	396	824
Unique Genes	4,410	3,735	5,369
Jaccard Coefficient	0.360	0.096	0.133
M. musculus	KEGG vs WikiPathways	WikiPathways vs MouseCyc	MouseCyc vs KEGG
Overlap Genes	2,611	532	919
Unique Genes	5,168	4,214	5,662
Jaccard Coefficient	0.336	0.112	0.140
S. cerevisiae	KEGG vs WikiPathways	WikiPathways vs YeastCyc	YeastCyc vs KEGG
Overlap Genes	801	402	480
Unique Genes	996	601	1,317
Jaccard Coefficient	0.446	0.400	0.267
M. tuberculosis H37Rv	KEGG vs WikiPathways	WikiPathways vs MTBRvCyc	MTBRvCyc vs KEGG
Overlap Genes	141	60	432
Unique Genes	948	525	707
Jaccard Coefficient	0.129	0.103	0.379

Table 3.3: Summary of overlapping gene proportions. Summary of the number of overlap genes, number of unique genes, and Jaccard coefficient among three representative databases.

pair number in each of the source databases can be found in Table 3.2. To calculate the agreement of genes and gene pairs in different databases, we obtain all the non-redundant genes and gene pairs (without considering the types of relationships) in different databases. We then calculate how many genes and gene pairs are common between two databases being compared. The Jaccard coefficient between two datasets being compared is calculated. Results are shown in the form of pie charts in Figure 3.1 and Figure 3.2, the detail statistics are listed in Table 3.3 and Table 3.4.

To analyze the agreement of the pathways in different databases, we only look at the pathway names in different databases, and calculate how many pathways two databases have in common. To find similar pathway names, we implement a "Longest Common Substring" algorithm. Our program can detect similar pathway names very accurately; detailed techniques will be explained in the following section. In this analysis we only search the related pathway between databases rather than within databases. The

H. sapiens	KEGG vs WikiPathways	WikiPathways vs HumanCyc	HumanCyc vs KEGG
Overlap Gene Pairs	1198	468	1,270
Unique Gene Pairs	45,205	23,060	32,987
Jaccard Coefficient	0.026	0.020	0.037
M. musculus	KEGG vs WikiPathways	WikiPathways vs MouseCyc	MouseCyc vs KEGG
Overlap Gene Pairs	875	1,242	2,068
Unique Gene Pairs	55,489	33,312	38,891
Jaccard Coefficient	0.016	0.036	0.050
S. cerevisiae	KEGG vs WikiPathways	WikiPathways vs YeastCyc	YeastCyc vs KEGG
Overlap Gene Pairs	35	9	419
Unique Gene Pairs	2,909	1,479	3,524
Jaccard Coefficient	0.012	0.006	0.106
M. tuberculosis H37Rv	KEGG vs WikiPathways	WikiPathways vs MTBRvCyc	MTBRvCyc vs KEGG
Overlap Gene Pairs	9	8	358
Unique Gene Pairs	3,819	2,810	5,823
Jaccard Coefficient	0.002	0.003	0.058

Table 3.4: Summary of overlapping gene pair proportions. Summary of the number of overlap gene pairs, number of unique gene pairs, and Jaccard coefficient among three representative databases.

results are presented in Figure 3.3.

The two experiments above are analyses at the database level. Next, to analyze—at the pathway level—the agreement of genes and gene pairs of the same pathway in different databases, We calculate the overlap of the genes and gene pairs in the chosen pathway in different databases. The results are summarized in Table 3.5.

3.3.3 Integration of pathway-gene and pathway-gene pair relationships

From the analyses above, we realize the lack of comprehensiveness and consistency of different pathway databases at both the database level and the pathway level. Hence, we should use the integrated information from all the databases rather than rely on any single source. The inconsistent referrals to pathway names further strengthen the necessity of integrating the pathway-gene and pathway-gene pair relationships from

M. musculus	TCA cycle pathway	KEGG vs WikiPathways	KEGG vs MouseCyc	MouseCyc vs WikiPathways
Gene	Count	31 vs 30	31 vs 13	13 vs 30
	Overlap	24	13	11
	Jaccard Coefficient	0.65	0.42	0.34
Gene Pair	Count	100 vs 30	100 vs 24	24 vs 30
	Overlap	10	9	7
	Jaccard Coefficient	0.083	0.078	0.149

H. sapiens	Fatty Acid Biosynthesis	KEGG vs WikiPathways	KEGG vs HumanCyc	HumanCyc vs WikiPathways
Gene	Count	6 vs 22	6 vs 2	2 vs 22
	Overlap	3	2	1
	Jaccard Coefficient	0.12	0.33	0.04
Gene Pair	Count	12 vs 29	12 vs 2	2 vs 29
	Overlap	1	1	0
	Jaccard Coefficient	0.025	0.077	0.0

M. tuberculosis H37Rv	TCA cycle pathway	KEGG vs WikiPathways	KEGG vs MTBRvCyc	MTBRvCyc vs WikiPathways
Gene	Count	35 vs 34	35 vs 10	10 vs 34
	Overlap	34	10	10
	Jaccard Coefficient	0.97	0.29	0.29
Gene Pair	Count	107 vs 37	107 vs 19	19 vs 37
	Overlap	3	9	5
	Jaccard Coefficient	0.021	0.077	0.098

Table 3.5: Table showing data overlap for same chosen pathways in difference source databases. This table shows the calculation of gene/gene pair differences and overlap between the different source databases for the same chosen pathways.

different databases into one unified and comprehensive information source.

To find all related pathways both among and within databases which have inconsistent referrals to pathway names (both name variations and different levels of emphases), we implement a refinement of the Longest Common Substring (LCS) algorithm to identify related pathway names. LCS was shown by Soh et al. (2010) to be superior for identifying related pathways in different databases, compared to approaches based on large overlap of genes and interacting gene pairs.

The common LCS algorithm based on dynamic programming works like this: when comparing two strings, the more similar they are, the higher alignment score they have. In our program, the alignment score is the number of aligned characters. We also compute the alignment ratio, which is two times the alignment score divided by the sum of the length of the two strings. To identify the related pathway names in two databases' pathway name lists (x and y), we iterate each name in the list x and search against all the names in the list y; for each name in x, we report the best hit in y. "Best hit" means that, for each name from x, when searched against all the names in y, the one that gets the highest alignment ratio is reported as the best hit for this round. We do not use the alignment score to report best hits because the alignment ratio proves to perform better. This is because some related pathway names do not have very high alignment scores due to the short length of two strings, but the similarities of the two strings can be revealed accurately by the high alignment ratio when compared to other not-so-similar long strings in a single round of search. For example, suppose the name Xa is very short and is searched against the name list y. Suppose there is a very similar short name Ya which aligns all characters in Xa except one character. Suppose there is also a very different but long name Yb which aligns all characters in Xa. It is obvious that the alignment score $Xa - Ya$ is lower than the alignment score $Xa - Yb$, while the alignment ratio of $Xa - Ya$ is the higher of the two. Thus using the alignment score to report the best hit is not as good as using the alignment ratio.

From many background experiments, we realize that relying only on best hits can result in some noise, since many pathway names in the list x may not have any related pathway names in the list y. Our strategy is to introduce more stringent requirement to increase the precision of the reported best hits. We require that either of the following two additional (empirically determined) conditions to be satisfied:

1. Alignment score > the length of shorter string -1 & alignment ratio $>= 0.5$, or

2. Alignment ratio > 0.91

Combined with this additional requirement, our program achieves high precision and recall in identifying related pathway names. Nevertheless, a small number of pathways which do not describe the same pathway, but have very similar names, are still incorrectly identified by the methods described above as related pathways. "VEGF signaling pathway" and "EGFR1 Signaling Pathway", "T Cell Receptor Signaling Pathway" and "B Cell Receptor Signaling Pathway", etc. are examples of this kind of mismatches. Our approach to solve this problem is by using a "error-prone words pair list" to filter potential mismatches. For example, if in a candidate related pathway pair, one pathway name has one partner of an "error-prone words pair"(EGFR1) and the other pathway name contains the other partner in the "error-prone words pair"(VEGF), this pair of candidate related pathways is discarded by our program. This approach successfully gets rid of mismatched pathways without compromising the identification of related pathways. Although a little manual curation is needed for initializing the "error-prone words pair list", the curation work load is much less after the first time, since only a few changes or supplementations of "error-prone words pair list" are needed when processing different groups of pathway names. Moreover, it is suitable for many different pathways in different organisms.

We run our program to compare pathway names within each database and between the databases. After obtaining all the related pathways, our program uses a disjoint

IntPath	KEGG	WikiPathways	MouseCyc
Fatty Acid Biosynthesis	Fatty acid biosynthesis	Fatty Acid Biosynthesis	1. fatty acid biosynthesis initiation II 2. very long chain fatty acid biosynthesis 3. fatty acid biosynthesis initiation III
Cholesterol Biosynthesis		Cholesterol Biosynthesis	1. cholesterol biosynthesis III (via desmosterol) 2. cholesterol biosynthesis II (via 24,25-dihydro-lanosterol) 3. cholesterol biosynthesis I 4. superpathway of cholesterol biosynthesis
TCA cycle	Citrate cycle (TCA cycle)	TCA cycle	TCA Cycle
Glycolysis and Gluconeogenesis	Glycolysis / Gluconeogenesis	Glycolysis and Gluconeogenesis	1. glycolysis I 2. glycolysis II

Table 3.6: Examples of inconsistent referrals to pathway names in *M. musculus*. The table shows several examples of the same pathways with inconsistent referrals to pathway names in different databases.

set data structure to store all the identified related pathways and then groups together all the related pathways under a general pathway name. The general pathway name is chosen as the shortest pathway names from among the identified related pathways. The shortest pathway name is usually suitable to be the name of the integrated pathway. However, in some cases, the shortest name contains "suffix" or "prefix"—like "I", "II"—that causes the integrated pathway name to give the wrong idea of describing only a specific aspect of the integrated pathway. So our program removes such suffixes and prefixes when generating integrated pathway names. In addition, there are also a small number of cases where several similar pathways are included in one pathway name—an example is shown in the last row of Table 3.6.

In these cases, the shortest name is not appropriate as the name of the integrated pathway. For these small number of cases, we replace the keyword of the integrated pathway name to cover more pathway information. After all the processing steps described above, we can be sure that the integrated pathway names in IntPath is

H. sapiens	KEGG	HumanCyc	WikiPathways
KEGG	5	3	29
HumanCyc	3	34	12
WikiPathways	29	12	4
M. musculus	KEGG	MouseCyc	WikiPathways
KEGG	6	6	32
MouseCyc	6	61	14
WikiPathways	32	14	10
S. cerevisiae	KEGG	YeastCyc	WikiPathways
KEGG	1	10	11
YeastCyc	10	25	74
WikiPathways	11	74	15
M. tuberculosis H37Rv	KEGG	MTBRvCyc	WikiPathways
KEGG	1	7	8
MTBRvCyc	7	35	2
WikiPathways	8	2	0

Table 3.7: Number of related pathways. Summary of the number of identified related pathways within and among databases.

correct and accurate. The numbers of identified related pathway names are listed in Table 3.7.

The number of pathways, average number of genes per pathway, and average number of gene pairs per pathway in each database, before and after this integration, is given in Table 3.8.

3.3.4 IntPath web interface and web service

IntPath is developed using JSP and MySQL. The web service is created and published using AXIS2.

3.4 Results

3.4.1 Extraction and normalization of pathway-gene and pathway-gene pair relationships

In order to overcome the limitation of incompatible data formats, we directly extract from the XML files (KGML, GPLM, BioPAX) of each pathway database and obtain

H. sapiens	No. of Pathways BEFORE integration	Average No. of genes/pathway	Average No. of gene pairs/pathway
WikiPathways	135 pathways	46.3	166.2
HumanCyc	290 pathways	7.2	33.0
KEGG	237 pathways	72.4	171.3
H. sapiens	No. of unique Pathways AFTER integration	Average No. of genes/pathway	Average No. of gene pairs/pathway
WikiPathways	100 pathways	42.7	157.4
HumanCyc	225 pathways	7.2	31.6
KEGG	201 pathways	72.6	165.3
Integrated Pathways	57 pathways	59.5	263.6

M. musculus	No. of Pathways BEFORE integration	Average No. of genes/pathway	Average No. of gene pairs/pathway
WikiPathways	140 pathways	57.8	209.1
MouseCyc	323 pathways	8.0	61.4
KEGG	218 pathways	74.6	194.8
M. musculus	No. of unique Pathways AFTER integration	Average No. of genes/pathway	Average No. of gene pairs/pathway
WikiPathways	97 pathways	56.8	242.8
MouseCyc	204 pathways	7.4	43.0
KEGG	172 pathways	77.9	187.3
Integrated Pathways	85 pathways	52.6	260.9

S. cerevisiae	No. of Pathways BEFORE integration	Average No. of genes/pathway	Average No. of gene pairs/pathway
WikiPathways	125 pathways	11.8	0.5
YeastCyc	184 pathways	6.5	13.4
KEGG	98 pathways	35.2	34.7
S. cerevisiae	No. of unique Pathways AFTER integration	Average No. of genes/pathway	Average No. of gene pairs/pathway
WikiPathways	45 pathways	15.1	0.2
YeastCyc	85 pathways	5.8	11.6
KEGG	80 pathways	38.0	35.0
Integrated Pathways	76 pathways	14.1	25.2

M. tuberculosis H37Rv	No. of Pathways BEFORE integration	Average No. of genes/pathway	Average No. of gene pairs/pathway
WikiPathways	8 pathways	22.3	7.8
MTBRvCyc	234 pathways	5.7	18.9
KEGG	110 pathways	32.5	47.5
M. tuberculosis H37Rv	No. of unique Pathways AFTER integration	Average No. of genes/pathway	Average No. of gene pairs/pathway
WikiPathways	0 pathways		
MTBRvCyc	171 pathways	5.9	21.0
KEGG	94 pathways	35.4	51.7
Integrated Pathways	35 pathways	12.3	25.4

Table 3.8: Summary of number of pathways, average number of genes per pathway and average number of gene pairs per pathway before and after integration. The table below shows the number of pathways from major pathway databases before and after integration.

the gene relationships. To deal with inconsistent molecular representations, we normalize the gene representations into a unified gene ID. The IntPath unified gene ID (which adopts a set of the most commonly used gene names) is compatible with the gene names used in most public repositories. A summary of the number of pathways, genes and gene pairs from different databases after normalization is given in Table 3.2. To tackle inconsistent molecular relationship representations, we also normalized the relationships of different databases into the IntPath unified relationship types as shown in Table 3.1.

3.4.2 Evaluation of normalized pathway genes and gene pairs from different databases

After obtaining the normalized pathway-gene and pathway-gene pair relationships, we are able to analyze the comprehensiveness and agreement among the different pathway databases on different aspects.

The results from analyzing the overlap of genes and gene pairs in different databases are presented in pie charts in Figures 3.1 and 3.2. The detailed statistics are summarized in Tables 3.3 and 3.4. These results prove that the overlap of genes and gene pairs in different databases are very low. This result is in accord with similar experiments done on human pathway databases (Soh et al., 2010).

From the results on the overlap of the pathways in different databases we can see there is also a strikingly low overlap of pathways among the different databases; see Figures 3.3. This demonstrates the obvious low level of comprehensiveness in the databases analyzed, also in accord with the experiments on human pathway databases described in (Soh et al., 2010).

Zooming in from the database level to the individual pathway level, we analyze the agreement of genes and gene pairs of the same pathway in different databases. The results are listed in Table 3.5. The agreement of different databases at the pathway

level is also not as high as we expected (especially for gene pairs), which proves the low level of consistency between these databases on the same pathway.

The comparative analyses from the above three aspects clearly exhibit the incomprehensiveness and inconsistency among the pathway databases. This suggests that the integration of the extracted and normalized information from different databases into a unified and comprehensive resource is very necessary.

3.4.3 Integration of pathway-gene and pathway-gene pair relationships

The results above demonstrate that relying only on a single source of pathway information from any of the databases is not reasonable. Moreover, we have also discovered the problem of inconsistent referrals to pathway names. Table 3.6 lists some examples of the same pathway under inconsistent names in different databases. Those are just a few typical examples; there are many pathways with similar situations which need to be properly addressed. Therefore, it is of great necessity to integrate all the pathway-gene and pathway-gene pair relationships from different databases into a comprehensive and unified source.

In the integrated pathways, all the related pathways with inconsistent names should be merged. (i) The inconsistent referrals to pathway names are partially caused by the different levels of emphases on the same pathway in different databases. One database (BioCyc) may emphasize on some very specific aspects of a certain large pathway; so this large pathway is broken up in this database into different pathways with similar/related names, yet all describing the detailed aspects of the original large pathway; see Table 3.6.

However, the other two databases may emphasize on a more general level and, therefore only use a general and often shorter pathway name. When merging pathways from different databases into integrated pathways, we should unify the different levels

of emphases. We decide to choose a more general level rather than a detailed level. (ii) When merging the same pathways with different levels of emphases in different databases, if we have already merged one detailed-level pathway into a general-level pathway, all other related detailed-level pathways in the databases should be merged into this general-level pathway. After merging all the related pathways we should use a general pathway name (usually the shortest one) to represent the integrated pathway. (iii) The distinct differences between our integrated pathway gene relationships and conventional pictorial pathway map indicate a more general level is suitable. We are primarily focusing on gene relationships, but not on other the relationships in the pathways (protein-compound relationships, compound-compound relationships, and so on.) in this version of IntPath. This emphasis results in less enthusiasm on the detailed level of individual pathways, and we lack sufficient information (just gene relationships) to emphasize on the detailed level in most cases. (iv) The common problem of gene relationships is the sparseness in each pathway; and putting emphasis on the detailed aspect of certain pathway could render the data in a single pathway too sparse to be useful.

For the reasons listed above, we should merge all the related pathways under the same general name into one comprehensive pathway (among and within databases). And after merging, we should use the general pathway name which is usually the shortest name among all the comparing pathway names.

The results of identified related pathways both within and among databases are summarized in Table 3.7. From the number of related pathways within databases, we find BioCyc and MouseCyc emphasize more on the detailed aspect of pathways; therefore, more related pathway names are identified. In IntPath, all the related pathways within and among databases are grouped together with the integrated pathway name. The number of pathways, average number of genes per pathway and average number of gene pairs per pathway, before and after integration, in the four IntPath included

organisms are given in Table 3.8. The statistics listed in Table 3.8 clearly show that in integrated pathways there is a significant increase of average node degree (average node degree = average no. of gene pairs per pathway / average no. of genes per pathway), which means significant increase of gene relationships of each gene on average in the integrated pathways. There is also a considerable increase of average no. of gene pairs per pathway in the integrated pathways, which indicates richer gene relationships on average in each pathway. In some sense, the integration approach partially solves the sparseness of pathway-gene relationships in MouseCyc and BioCyc.

We have accomplished in IntPath the integration of pathway-gene and pathway-gene pair relationships, achieving compatible data formats, consistent molecular representations, consistent relationship representations, consistent referrals to pathway names and comprehensive data.

3.4.4 IntPath web interface and web service

The web interface of IntPath comprises the following parts: Home, Gene List Analysis Tools (Identify Pathways and Analyze Distances), API Toolkit, Statistics, Tutorial, and Download. In order to facilitate convenient access of IntPath data through local programs, the API functions are also supported by IntPath web service. An overview of the IntPath system is shown in Figure 3.4. The core functions of IntPath are represented in Figure 3.5. An explanation of each part is given below.

Home: It is to introduce the objective of IntPath, what the major contribution of this database is and what the specific problems that we wish to solve through this database are. We also indicate the analysis tools supported in this database, the publications related to these analysis tools, and which species are currently included in our database. This Home page of IntPath is a summary of the general information of the database.

Identify Pathways: The function of “Identify Pathways” uses the hyper-geometric

test to find the most significant pathways given an input gene list. Through this tool, users can have a clear insight of which pathway is most related to the input gene list. For each result returned, details like p-value are also given.

Analyze Distances: The function of "Analyze Distances" is to tell the similarities between the two input gene lists from a pathway perspective. To perform the distance analysis, first the hyper-geometric test is used to find the most significant pathways of the two input gene lists, then the Floyd-Warshall algorithm is used to calculate the "distances" between the two pathways. STRING PPI datasets (version 9.0) is used in the distance calculation between two pathways in the current version of IntPath (V2.0). The "distances" provide a reference in telling the relationships between two specific pathways, and it can be very useful, e.g., in identifying how "far" it will take for a normal pathway to transform into a diseased pathway. For a detailed explanation of "Analyze Distances" and its application in biomedical research, please refer to methods described in (Goh et al., 2011).

Statistics: This statistics section gives users an overall insight of IntPath. Users can easily get the following statistics: number of genes, number of gene pairs, number of integrated pathways, number of original KEGG pathways, number of original WikiPathways pathways, number of original BioCyc(MouseCyc) pathways, and number of source databases. The default option is "All statistics" which displays all the statistics listed above.

API Toolkit : We provide powerful as well as flexible API functions of our IntPath database. Users can both call the API functions using their local programs through IntPath's web service or using API functions by directly retrieving information through IntPath's web interface. The following API functions are supported, getGeneID, getDBPathways, getPathway, getPathwayGenes, getGenePathways, getPathwayInteractions, getPathwayDifference, getIntPathGenes, getIntPathGenePairs and getIntPathPathways. The explanation and user guide of each API function can be found in the

Tutorial page.

Download: Some users may have other requirements of data analyses that are not met by IntPath in the current version. Some users may also have different application purposes of IntPath. To cope with a variety of needs, we release all our IntPath data in this "Download" section, where users can obtain all IntPath data in two different formats: (1) text format (*.txt), this compressed package includes three text files, (a) the integrated pathway-gene relationships, (b) the integrated pathway-gene pair relationships and (c) the normalized group-genes list; and (2) sqldump format (*.sql), which is based on the integrated data we have prepared and stored in 6 tables in each sqldump (each organism is a separate sqldump).

3.5 Discussion

3.5.1 Comments on WikiPathways

The "wiki-style" of WikiPathways makes this database more casual than other databases. It is good for the community to freely maintain and share knowledge through WikiPathways. On the other hand, it causes many problems for automatic information retrieval. One of the limitations is the slight inconsistency among the formats of GPML as mentioned before—some key tags can be upper or lower cases. GPML is more different from other XML formats. GPML emphasizes more on pictorial information; therefore, most of the objects on the file are more likely to be recorded for their positional information. Worse, some GPML files even do not have a "graphID" record; and for these GPML files, whole information of certain pathways is given by the positional information on the pathway map. For these GPML files, judging the relationships between two genes is solely dependent on the positional information. It may be easy for the human eye to look at the pictorial format of the pathway map; but it is hard for computer programs to retrieve accurate information automatically. Attempts on spatial clustering

have been made. But these attempts also introduce a substantial amount of noise. Therefore we decide to discard this noisy information at the current stage.

Recently, WikiPathways begins to support web service and BioPAX. We have tried solving the problem mentioned above using WikiPathways web service and directly extracting from BioPAX format; but no improvement has been achieved.

Web service has not solved the problem of those GPML files that do not have "graphID" record. For example, our program fails to extract reliable gene relationships from the pathway "Mm Androgen Receptor Signaling Pathway WP252_35669" by calling the WikiPathways API function "findInteractions". It is supposed to find interactions defined in WikiPathways' pathways. In our experiments, it works in finding interactions in other pathways. Extensive experiments have been made using different ways to call the "findInteractions" function. Yet nothing related to the WP252 pathway is returned. On the png graph we can see there are lots of interactions in this WP252 pathway. These kinds of experiments have been attempted many times on several pathways. All have failed to find the "interactions" or gene relationships in the specific pathways that lack "graphID" entry.

We turn to BioPAX files which have recently been supported on WikiPathways for a solution. We specifically run our program on the pathway BioPAX files whose corresponding GPML files do not have "graphID" records. Our program successfully retrieved gene relationships from BioPAX files in BioCyc and MouseCyc, but not on these specific BioPAX files in WikiPathways (for example, "Sc Cell Cycle and Cell Division WP414_21554" and "Sc Glycolysis and Gluconeogenesis WP515_42806"). We also try to visualize those specific BioPAX files on Cytoscape; but no relationship can be visualized from these files.

The gene ID problems in WikiPathways is also quite serious. There are two places to retrieve gene ID information from the GPML "</DataNode>" entry, one is from "TextLabel" and the other is from "<Xref Database" IDs. Usually gene IDs in "Text-

Label" are gene symbol, while gene IDs in "<Xref Database" can be the gene IDs from different public databases, like Entrez, Ensembl, UniProt, and so on. Getting gene ID information from both of these two fields is necessary. It is not uncommon for the WikiPathways database to have errors and problems in both fields. In most cases, erroneous gene IDs from "TextLabel" also do not have any information in "<Xref Database". The erroneous gene IDs can be gene symbols or EC numbers that cannot be found in the target organism to which the pathway map belongs; they can also be common gene names without any information in "<Xref Database", or they can be just upper- or lower-case flaws. In our program, both information from the two fields,"TextLabel" and "<Xref Database" are retrieved. For gene IDs where information from both of these fields are problematic, manual curation is adopted to deal with them, generally by removing them from IntPath.

3.5.2 Access, update and extension of IntPath

IntPath and all its data have been released online at `http://compbio.ddns.comp.nus.edu.sg:8080/IntPath`. As some analyses in Chapter 2, Chapter 4, Chapter 5 already used IntPath data, we believe our work here can facilitate a variety of works that need to refer to pathway information.

IntPath heavily depends on source pathway data from all the pathway databases and most databases are updated quite frequently. The important question is: Can we keep our data updated in a timely fashion? The answer is: Yes.

The "IntPath Data Preparation" program is streamlined and automated in performing the extraction, normalization, integration processes and directly outputing into MySQL databases and text files. For organism already included in IntPath, running the program for each update takes a short time; and we will maintain a regular update of IntPath in the long term. Another key question is whether we can extend our approach to other organisms. Currently, we have already included four organisms—*S.*

cerevisiae, *M. tuberculosis* H37Rv, *H. Sapiens* and *M. musculus*—and we will include more in future releases of IntPath. Extending the methodology to include other organisms just needs modifying the regular expressions for extracting GPML and KGML files; preparing the gene ID mapping files; manually correcting some possible errors of the gene IDs introduced by the source databases (like WikiPathways gene ID problems) and, if necessary, updating the "error-prone words pair list"; and reviewing integrated pathway names. Therefore, the whole process of including other organisms in IntPath takes a short time. We will include more model organisms and important pathogens in IntPath in future releases.

3.5.3 Outlook of IntPath

Pathway data have wide application in a variety of studies(Wong, 2011), for example, they are used to improve gene expression profile analysis(Soh et al., 2011), to make protein function prediction in the absence of sequence homology(Hawkins and Kihara, 2007), to better identify targets for disrupting pathogen drug resistance mechanisms(Wong and Liu, 2010), to better identify disease genes(Li and Agarwal, 2009), to improve robustness of proteomic profile analysis(Goh et al., 2013), and drug target prioritization(Yadav et al., 2013). We believe IntPath will contribute greatly to all these aspects of application.

In the near future, more functions and analysis tools will be supported in IntPath—for example, clustering algorithms for microarray studies using the IntPath data as background knowledge, visualization tools of interaction and relationship, more powerful algorithms to identify pathways given user-specified input gene lists, and more API functions. Moreover, in this version of IntPath we only take gene relationships into account; in a future version, IntPath will also consider other important relationships in the pathways—like protein-compound relationships, compound-compound relationships, and so on. Meanwhile, in future releases, more organisms will be included.

We wish our continuing effort can make IntPath one of the most useful databases in pathway studies that can benefit a variety of related researches.

3.6 Conclusion

The five limitations of current pathway databases that hamper effective use of pathway information have been overcome in this Chapter. We solve the problem of incompatible data formats in different databases by extracting the pathway-gene and pathway-gene pair relationships. The limitations of inconsistent molecular representations and inconsistent molecular relationship representations have been overcome by our normalization of the data into common gene name representations and common relationship types which are compatible with other database. The problems of inconsistent referrals to pathway names and incomprehensive data from different databases have been solved by the integration of pathway-gene and pathway-gene pair relationships into a unified and comprehensive data source.

We achieve compatible data formats, consistent molecular representations, consistent relationship representations, consistent referrals to pathway names and comprehensive data in our IntPath database for several organisms—viz., *H. sapiens*, *S. cerevisiae*, *M. musculus* and *M. tuberculosis* H37Rv. IntPath can maintain a regular update in these organisms and, the methodology we describe here can be applied to other organisms straightforwardly.

We believe IntPath will not only facilitate convenient access of the integrated pathway gene relationship data for model organisms and important pathogens but also greatly boost data analysis and application to many related studies through the analysis tools and API functions provided in the database.

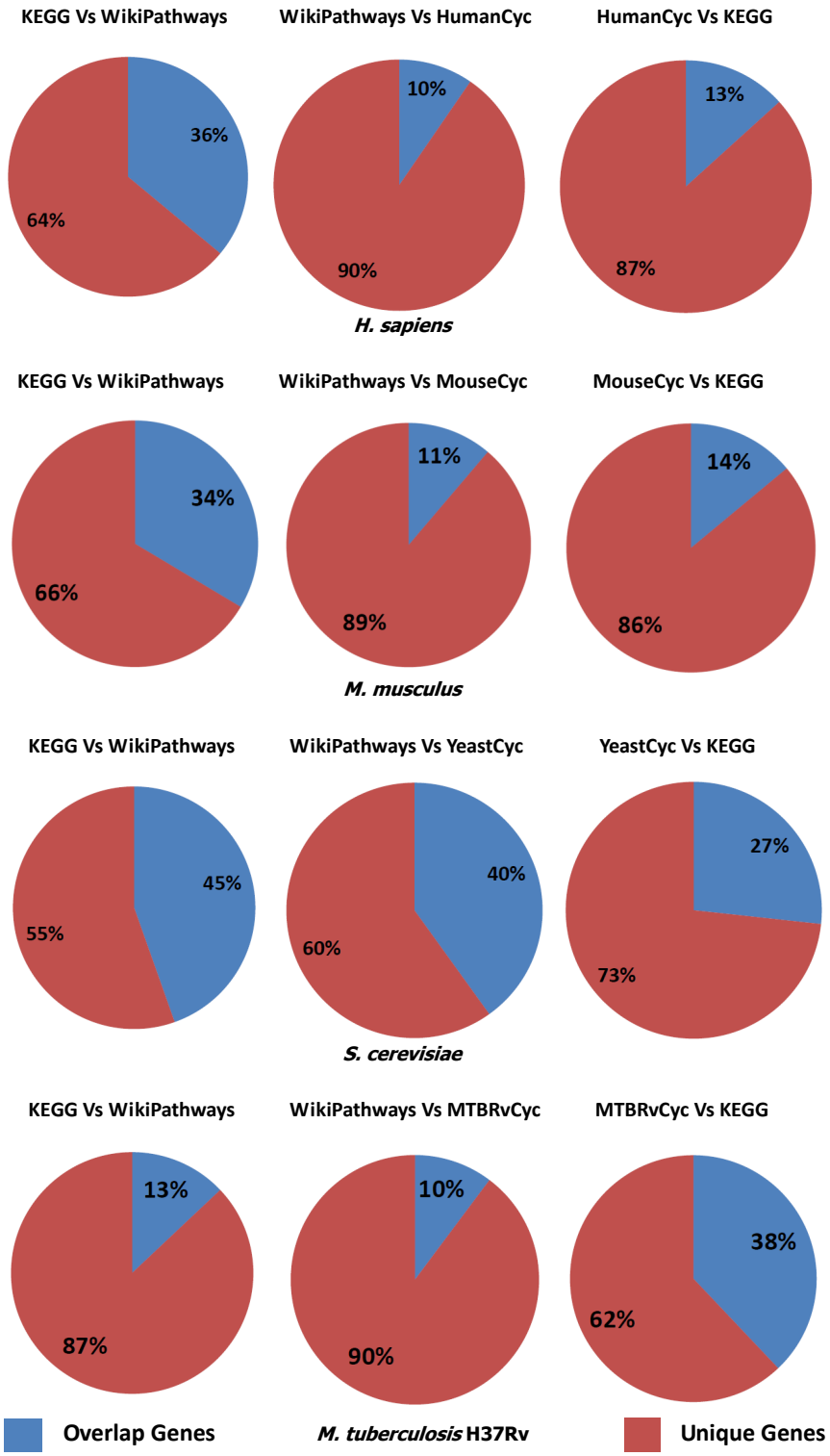


Figure 3.1: Pie charts depicting overlapping gene proportions. The red part refers to the proportions of unique genes while the blue part refers to proportions where there is an overlap of genes.

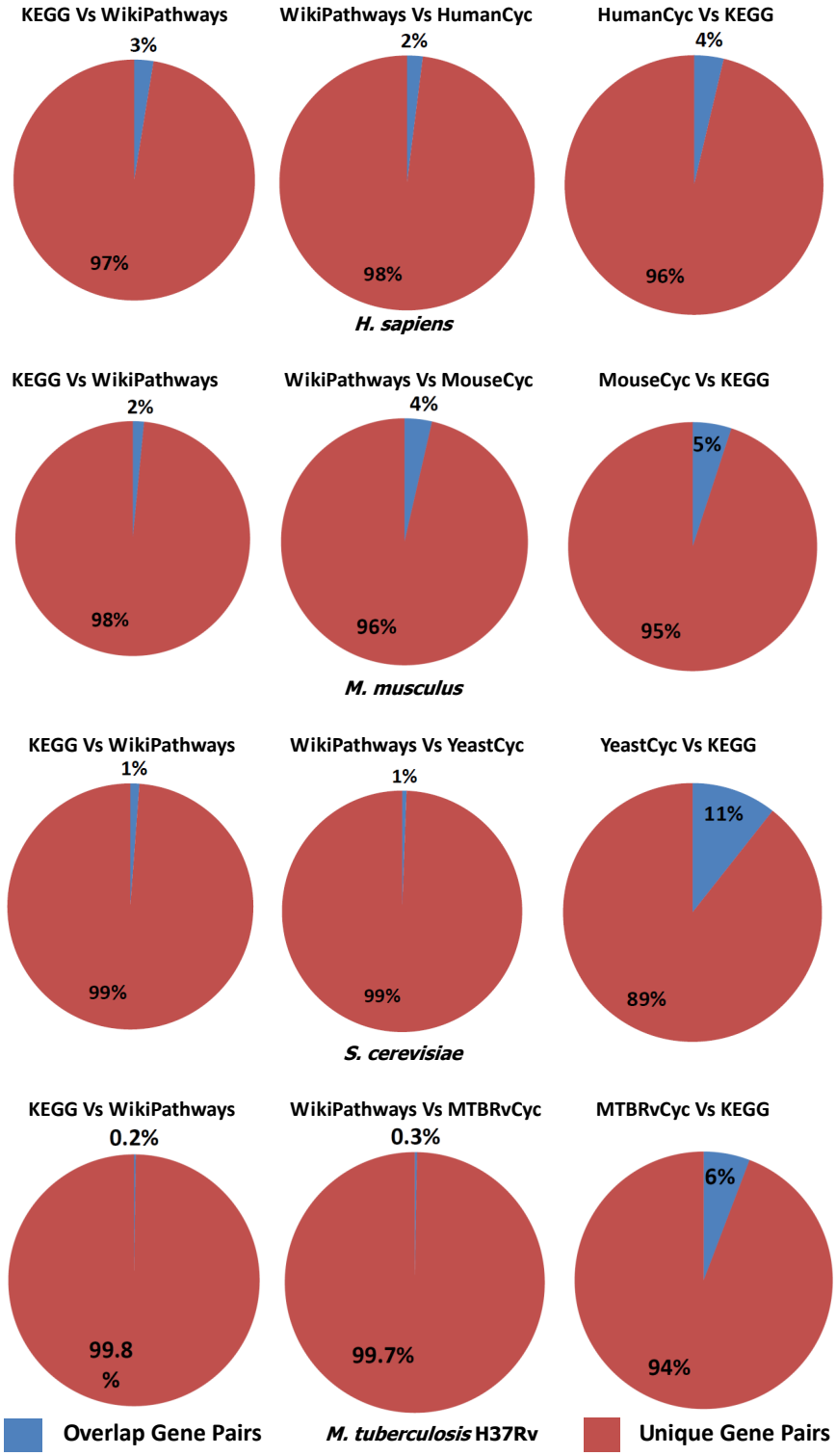


Figure 3.2: Pie charts depicting overlapping gene pair proportions. The red part refers to the proportions of unique gene pairs while the blue part refers to proportions where there is an overlap of gene pairs.

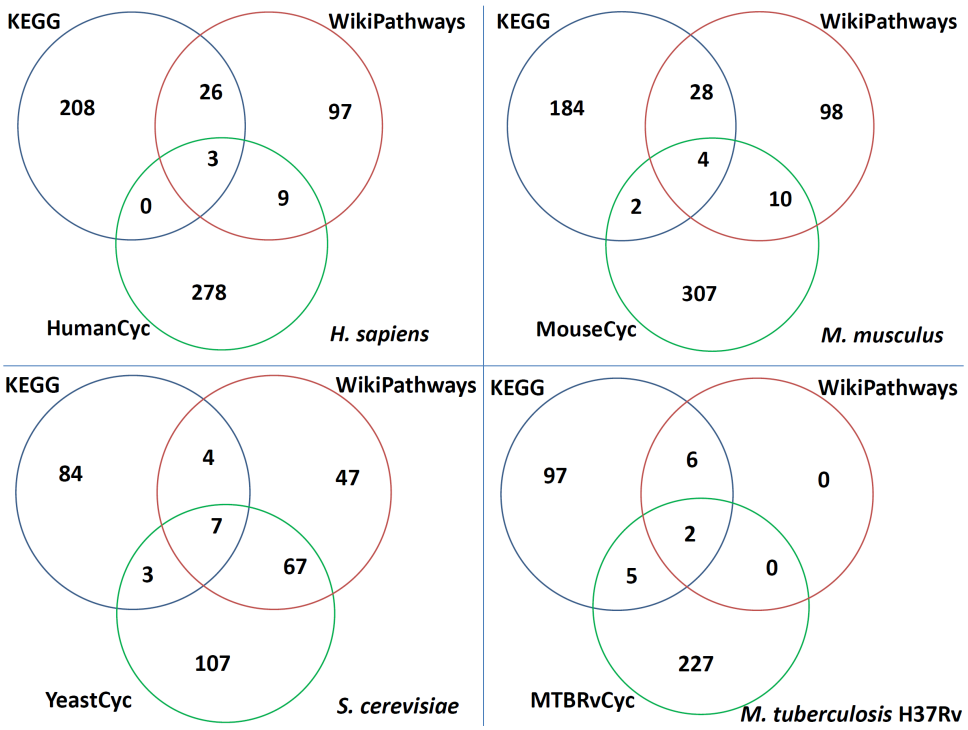


Figure 3.3: Venn diagram of pathways in different databases. Venn diagram depicting overlapping pathways across the three databases.

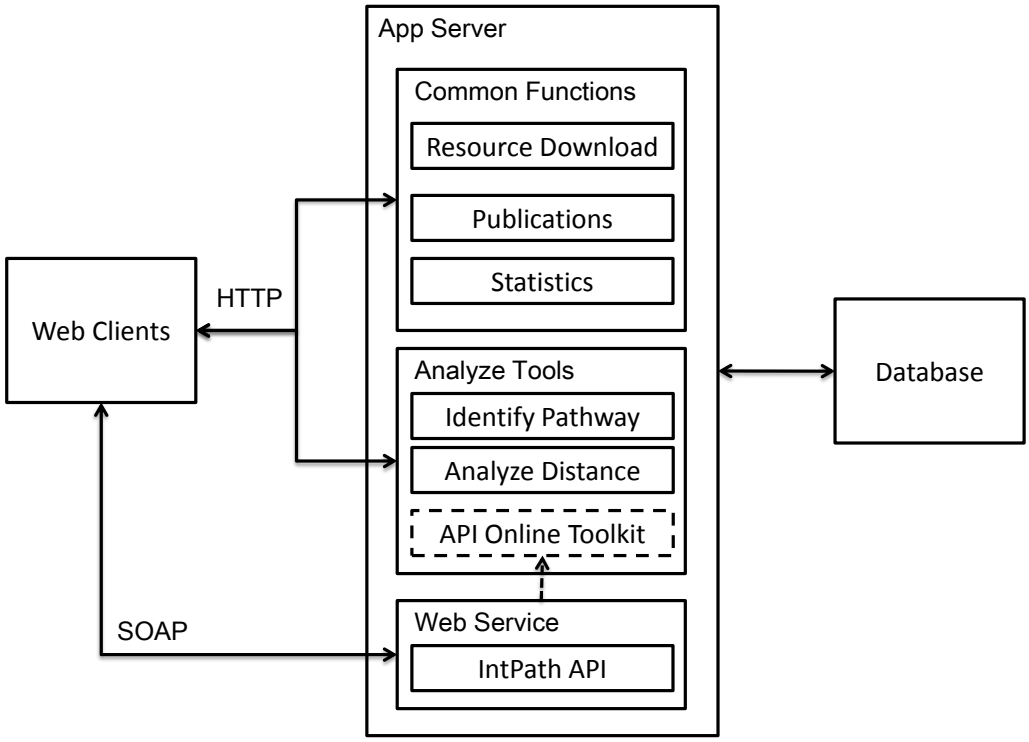


Figure 3.4: IntPath system overview. This figure shows the components of IntPath database, the relationships between those components and a clear indication on which components are supported by web service and which are supported by web interface.

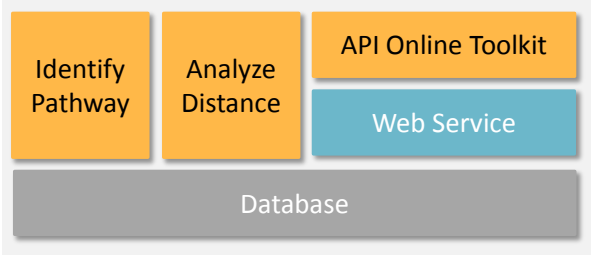


Figure 3.5: Core functions of IntPath. This figure shows the core functions of IntPath, the relationships between those core functions, database and web service.

Chapter 4

Stringent DDI-based Prediction

Domain-domain interaction (DDI) based prediction is one of the frequently used computational approaches in predicting both intra-species and inter-species PPIs. However, the performance of DDI-based host–pathogen PPI prediction has been rather limited.

We develop a stringent DDI-based prediction approach with emphasis on (i) differences between the specific domain sequences on annotated regions of proteins under the same domain ID and (ii) calculation of the interaction strength of predicted PPIs based on the interacting residues in their interaction interfaces. As long as the two amino acids have one of the atomic interactions (hydrogen bonds, electrostatic or van de Waals interactions) between two domain instances, they are defined as interacting residues in this study.

The stringent DDI-based prediction approach reported in this Chapter provides an accurate strategy for predicting host–pathogen PPIs. It also performs better than a conventional DDI-based approach in predicting PPIs. We have predicted a small set of accurate *H. sapiens–M. tuberculosis* H37Rv PPIs which could be very useful for a variety of related studies.

4.1 Background

Tuberculosis is an infectious disease which causes millions of deaths each year. *M. tuberculosis*—the causative agent of tuberculosis— infects around one-third of the world's population(Butler, 2000; Koul et al., 2004). Tuberculosis is one of the most common opportunistic infection in HIV-infected patients and it is also one of the most common death causes among HIV patients(Hestvik et al., 2006; Global Tuberculosis Programme, 2010).

Host–pathogen PPIs are essential for a pathogen's colonization, adhesion and invasion of host cells, which are crucial for the understanding of infection mechanism and the interaction between pathogen and host. Unfortunately, high-quality large-scale experimental host–pathogen PPIs are not available in many host–pathogen systems, especially between *H. sapiens* and *M. tuberculosis* H37Rv. Many computational approaches have been developed to predict host–pathogen PPIs including approaches based on homology, interacting domain/motif, structure, and even machine learning(Zhou et al., 2013). DDI-based approaches are often used for predicting both intra-species and inter-species PPIs, with the assumption that domain-domain interactions mediate the protein-protein interactions, because domains are the basic building blocks determining the structure and function of proteins(Zhou et al., 2013).

In this Chapter, we develop a stringent DDI-based approach for predicting the *H. sapiens–M. tuberculosis* H37Rv PPIs by taking into account of the differences between each specific domain sequence (we name it "domain instance") on each annotated region of proteins under the same domain ID. The interactions between query domain instances are made based on very stringent sequence alignment to the structural template domain instances. Moreover, we adopt an effective scoring strategy in ranking how likely the predicted proteins are interacting with each other by examining the interacting residues in the interaction interfaces. Thus, we are standing on a much more accurate and finer level of domain interaction by examining not only the sequence simi-

larity of each domain instances but also the interaction interface compatibility between them. In contrast, conventional DDI-based approaches generally use some popular tools to annotate the domains in proteins and then see whether two proteins contain a pair of domains whose IDs match a pair of domains that are known to interact in some other pair of proteins. Matching query domain instance to template domain instance based on domain ID—as done in such conventional DDI-based approaches—is rather coarse and often leads to matching of domain instances that do not have the same interaction interfaces.

Using gold standard *H. sapiens* PPIs, we assess the performance of our stringent DDI-based approach and the conventional DDI-based approach by comparing their precision-recall curves and the number of predicted PPIs overlapping with gold standard PPIs. We also use the percentage of coherent informative Gene Ontology(GO) annotations to assess the predicted *H. sapiens* PPIs to compare the performance of our stringent DDI-based approach and the conventional DDI-based approach. These assessments demonstrate that our stringent DDI-based approach has much better performance than a conventional DDI-based approach. Cellular compartment distribution analysis, pathway enrichment analysis, and functional category enrichment analysis supports the validity of our predicted *H. sapiens–M. tuberculosis* H37Rv PPI dataset. Our stringent DDI-based approach can be used for predicting host–pathogen PPIs in a variety of different host–pathogen systems. We have also discovered some interesting properties of both pathogen and host proteins participating in host–pathogen PPIs, including the tendency to have more domains, and the domains on the proteins involved in host–pathogen PPIs tend to have much higher degrees.

4.2 Methods

Our stringent DDI-based approach predicts PPIs by inferring domain instance interactions from structural template domain instance interactions. We accurately align

query protein domain instances to template domain instances using a stringent threshold(length difference $\leq 20\%$ and sequence similarity $\geq 50\%$) and transfer the possible interactions between template structural domain instances to our query domain instances. We then predict the possible PPIs from interacting query domain instances. The structural domain instances are extracted from the 3did database(Stein et al., 2011). Each interacting query domain instance pair is scored according to the similarity of the interaction residues in the interaction interfaces, and the best query instance score is used to represent the interaction strength of the predicted PPI (how likely the two proteins in the PPI are interacting each with other). We predict both host–pathogen (*H. sapiens–M. tuberculosis* H37Rv) and intra-species (*H. sapiens*) PPIs in this Chapter. For a comparison study, we use a conventional DDI-based approach(Dyer et al., 2007) to predict possible intra-species (*H. sapiens*) PPIs. We assess our stringent DDI-based approach and the conventional approach using gold standard *H. sapiens* PPIs and by the percentage of the predicted PPIs that have coherent informative GO annotation. These assessments show that our stringent DDI-based approach has better performance in predicting PPIs than the conventional approach. Cellular compartment distribution analysis, pathway enrichment analysis, and functional enrichment analysis support our prediction results and show that the predicted PPIs correspond to the *M. tuberculosis* H37Rv infection process. We further analyze some of the basic domain properties of proteins involved in the host–pathogen PPIN, comparing with other proteins involved in intra-species PPIN, by examining the number of domains and domain interaction degrees.

4.2.1 PPI prediction—our stringent DDI-based approach

It is a reasonable assumption that an observed interaction between two domain instances can be used to infer the interaction of another domain instance pair, provided the two domain instance pairs are sufficiently similar as to preserve the relevant in-

teraction interfaces. Specifically, consider two protein domains A and B. Let A_i and B_i be two instances of domain A and B, respectively. Suppose we know that these two instances have a direct physical interaction (from the crystal structure of a protein complex). Given the observation of A_i and B_i, one could infer the interaction of another instance pair of A and B, A_j and B_j, by using a sequence similarity threshold between (A_i, B_i) and (A_j, B_j).

In general, conventional DDI-based approaches disregard the details of the interaction between these domain instances in the real 3D space—i.e., the interaction interface between the two instances—and thus effectively matches the domain instances based on name. In contrast, we formulate a stringent approach that emphasizes the similarity of the interaction interface of the domain instances. Specifically, we assign a positive prediction score on pairs with high interface residue similarity with respect to the observed interaction instances in the existing protein structural data.

The data on structural domain instances, including the interacting domain pair, the structural and sequence details of interacting domain instances, the interacting residues in the interaction interfaces are extracted from the 3did database(Stein et al., 2011). These individual domain instances with 3did structural data serve as "template domain instances", and pairs of interacting domain instances with 3did structural data serve as "template interacting domain instance pairs". The fasta sequences of all *H. sapiens* and *M. tuberculosis* H37Rv proteins are obtained from Uniprot(The UniProt Consortium, 2012). Their respective protein domain annotations are obtained from InterPro(Hunter et al., 2012), from which we collect the sequences of domain instances which have at least one template domain instance from 3did. These domain instances are named the "query domain instances". They are aligned to each of the template domain instances under the same domain ID using the MUSCLE alignment program(Edgar, 2004). Only query domain instances meeting the stringent threshold of length difference $\leq 20\%$ and sequence similarity$\geq 50\%$ are kept for the following analysis. For each pair (A_i, B_i)

of query domain instances that meets the stringent alignment threshold to a template interacting domain instance pair (A, B), we infer the interaction interface residues in (A_i, B_i) as the residues that are aligned to the interaction interface residues in (A, B). A score of this interaction interface of (A_i, B_i) is then computed by summing the BLOSUM62 substitution score(Henikoff and Henikoff, 1992) between the residues in this interaction interface and the corresponding residues in the interaction interface of (A, B) that they are aligned to. This score is defined as the "domain instance interaction strength". Query domain instances with multiple possible template instances are scored based on the template with the best domain instance interaction strength. For any possible pair of proteins, if they have a query domain instance pair (one domain instance on each of the two proteins), then these two proteins are predicted to be interacting with an interaction score equaling the domain instance interaction strength of that query domain instance pair. If the protein pair has more than one underlying query domain instance interaction pair, then the query domain instance pair with the best score is used to represent the protein pair. This best score is taken as "interaction strength" of this protein pair.

We apply this DDI-based prediction approach on human proteins and this results in 839 predicted human intra-species PPIs. We also predict inter-species PPIs (*H. sapiens–M. tuberculosis* H37Rv) to identify a set of potential host–pathogen PPIs; the result is visualized in Figure 4.1.

4.2.2 PPI prediction—a convention DDI-based approach

The conventional DDI-based approach predicts how likely two proteins are interacting with each other by integrating known intra-species PPIs with domain profiles based on an association method (sequence-signature algorithm) proposed by Sprinzak and Margalit (2001) Specifically, domains are annotated in each protein in a known intra-species PPI dataset. Then, the probability $P(d, e)$ that two proteins containing a

specific pair of domains (d, e) would interact is estimated for each pair of domains in a Bayesian manner. Finally, given a new pair of proteins, their probability of interaction is estimated by a naive combination $(= 1 - \prod_i \prod_j (1 - P(d_i, e_j)))$ of the probabilities from each pair of domains (d_i, e_j) contained in the pair of proteins(Dyer et al., 2007). This predicted probability(called "interaction strength" of the conventional approach) can be used to rank the list of predicted PPIs.

This conventional DDI-based approach is applied to predict host–pathogen PPIs as follows. For each pair of proteins (one in *H. sapiens* and one in *M. tuberculosis*), we compute their probability of interactions as described above based on DDIs in a yeast physical PPI dataset collected from MINT(Zanzoni et al., 2002), BioGRID(Stark et al., 2011), and IntAct(Hermjakob et al., 2004). This conventional DDI-based approach is also applied to predict human intra-species PPIs. In this case, for each pair of proteins (both in *H. sapiens*), we compute their probability of interactions as described above based on DDIs in the same yeast physical PPI dataset. As a control study, we ensure that the domains considered are the same domain set considered in the stringent DDI-based approach—i.e., we restrict the domain set to domains contained in 3did.

4.2.3 Assessment based on gold standard *H. sapiens* PPIs

Because no large-scale high-quality *H. sapiens-M. tuberculosis* PPI dataset is currently available, we can only assess the performance of the stringent and the conventional DDI-based approaches in a intra-species system. We use both the stringent and the conventional DDI-based approach to predict possible *H. sapiens* PPIs and assess the predicted PPI datasets using gold standard *H. sapiens* PPIs. The gold standard *H. sapiens* PPIs are the physical PPIs collected from MINT(Zanzoni et al., 2002), BioGRID(Stark et al., 2011), and IntAct(Hermjakob et al., 2004). We sort the predicted *H. sapiens* PPIs according to their predicted "interaction strength" in the respective DDI-based approaches, and compare the top PPIs with the gold standard *H. sapiens*

PPIs. For the stringent DDI-based approach, we sort the prediction results and iterate 10 PPIs at a time—which means the first time we choose all the top 839 PPIs, the second time we choose the top 829 PPIs, etc.—and then we compare with the gold standard *H. sapiens* PPIs to calculate the precision and recall and plot the precision-recall curve. The precision-recall curve of the conventional DDI-based approach is plotted in the same way. The precision-recall curves are plotted together for a better comparison in Figure 4.2.

As the two PPI datasets predicted by the stringent and the conventional DDI-based approaches are very different in the number of PPIs, their precision-recall curves may not be sufficient for judging the performance of the two prediction approaches. So we choose some special points to provide a more informative statistics. The stringent DDI-based approach predicted 839 *H. sapiens* PPIs and 82 of which overlap with the gold standard PPIs. We consider a similar amount of conventional-approach predicted *H. sapiens* PPIs(top 885 PPIs), and see how many of these predicted PPIs overlap with gold standard. We also choose another point on the precision-recall curve, that has a similar number of overlapping PPIs with the gold standard as the stringent DDI-based approach, and see how many predictions are made by conventional DDI-based approach. The results are shown in Table 4.1.

4.2.4 Assessment using coherent informative GO annotation of predicted *H. sapiens* PPIs

A PPI is more likely to be real, if its two protein components have coherent GO annotation—i.e., the two proteins are annotated with at least one "informative" GO term in common. The percentage of PPIs having coherent GO annotation is also frequently used in assessing the quality of the PPI dataset(Zhou and Wong, 2011). Note that GO contains three hierarchical ontologies, and terms at the root level have more proteins annotated with them, while terms at the leaf level have fewer proteins

annotated with them. In order to avoid bias, we only keep informative GO terms for the assessment here. An informative GO term is defined as a GO term that has at least 30 proteins annotated with it but each of its child terms has fewer than 30 proteins annotated with it. This definition of informative GO term is also used in another work(Zhou and Wong, 2011) for assessing PPI dataset quality in *M. tuberculosis* H37Rv. For the PPI datasets predicted by the stringent DDI-based approach and by the conventional DDI-based approach, the PPIs in each dataset are sorted according to their respective "interaction strength" (which is an indicator of how likely the PPIs are real), then the percentage of PPIs that has coherent informative GO terms are calculated. For each dataset we move along from the bottom to the top to set the threshold of how many top PPIs are considered, and calculate the percentage of these PPIs having coherent informative GO terms. For the stringent DDI-based approach we choose an interval of 10 PPIs and move along from the bottom to the top(e.g. top 839 PPIs, top 829 PPIs, etc.), then calculate the percentage of PPIs that have coherent informative GO terms and plot the percentage; see Figure 4.3.

For the conventional DDI-based approach, we plot the percentage in the same way; but as the conventional DDI-based approach predicts much more PPIs, we choose interval of 1000 PPIs while making the plot; see Figure 4.4.

To better compare and assess the performance of the stringent and the conventional DDI-based approaches, we focus on the top 839 PPIs predicted by both approaches, choosing interval of 10 PPIs as we plot the percentage of PPIs having coherent GO annotation on the same figure; see Figure 4.5.

When assessing the quality of two PPI datasets based on informative GO terms, the number of GO terms that are annotated to the proteins of that PPI dataset also influences the percentage of PPIs having coherent informative GO terms in that dataset. Therefore we summarize the number of informative GO terms in the 839 PPIs predicted by the stringent DDI-based approach, and the number of informative GO terms in

the 724185 PPIs and in the top 839 PPIs predicted by the conventional DDI-based approach; see Table 4.2.

4.2.5 Cellular compartment distribution of *H. sapiens* proteins targeted by the predicted host–pathogen PPIs.

The assessments above prove that our stringent DDI-based approach has a much better performance than the conventional DDI-based approach in predicting more reliable intra-species PPIs. We next analyze the host–pathogen PPIs predicted by our stringent DDI-based approach.

The cellular compartments of the *H. sapiens* proteins targeted by the predicted *H. sapiens-M. tuberculosis* H37Rv PPIs are useful in telling the quality of the predicted host–pathogen PPIs. If the targeted *H. sapiens* proteins are located in cellular compartments that are very relevant to the pathogen's infection or are very likely to be involved in interactions with the pathogen, then the result supports the host–pathogen predictions. Gene Ontology (Cellular Compartment, CC) is a very comprehensive annotation system for human proteins. However, as the Gene Ontology is hierarchical, we only use informative CC terms for our analysis.

In contrast to using the coherent informative GO annotation for the assessment of the human intra-species PPI dataset, we choose a different resolution of the GO terms for the category distribution analysis of human proteins involved in *H. sapiens-M. tuberculosis* PPIs: An informative CC term is defined here to be a term that has at least 90 proteins annotated with it, but each of its child terms has less than 90 proteins annotate with it. The cellular compartment distribution tells how many proteins(and the percentage) in the datasets fall into each cellular compartment. We show the cellular compartments of the *H. sapiens* proteins that are targeted by the stringent DDI-based prediction approach in Table 4.3 and Figure 4.6.

4.2.6 Functional enrichment analysis of proteins involved in host–pathogen PPIs

Functional enrichment analysis is important for revealing the functional relevance of the proteins involved in the host–pathogen PPIs predicted by our stringent DDI-based approach. The presence of enriched(over-represented) functional categories that are closely related to pathogen infection, serves as a support for the validity of the predicted host–pathogen PPIs. The Gene Ontology (Molecular Function, MF) is a comprehensive functional annotation system. Therefore we conduct MF term enrichment analysis on the *H. sapiens* proteins involved in the *H. sapiens-M. tuberculosis* H37Rv PPIs predicted by our stringent DDI-based approach. We use the DAVID database (Dennis Jr et al., 2003) for the GO term enrichment analysis. Results are shown in Table 4.4 (significantly enriched level 5 MF terms, threshold "count > 2, p-value < 0.1").

On the other hand, as we have found in another work(Zhou and Wong, 2011), most of the GO annotations for *M. tuberculosis* H37Rv are not specific enough to provide effective functional enrichment analysis. Thus, the functional analysis of *M. tuberculosis* H37Rv proteins are not discussed in this Chapter.

4.2.7 Pathway enrichment analysis of proteins involved in host–pathogen PPIs

Pathway data are very important functional information for identifying a list of proteins' overall related functions in a cell. For a set of proteins which is significantly enriched in some pathways, it is very likely that this set of proteins play similar or co-ordinated roles *in vivo*. Thus, pathway enrichment analysis is also one of the most frequently used strategy for analyzing predicted host–pathogen PPIs.

We use the IntPath(Zhou et al., 2012) database for the pathway enrichment analysis. IntPath is currently one of the most comprehensive integrated pathway databases. The "Identify Pathways" function in IntPath can identify the pathway enrichment of an

input gene list. The "Identify Pathways" function in IntPath(Zhou et al., 2012) adopts the hypergeometric test to identify the input gene list's over-representation(enrichment) in the pathways. For the *H. sapiens* protein set predicted by the stringent DDI-based approach, the pathway enrichment analysis result is shown in Table 4.5.

We also analyze the pathway enrichments for the *M. tuberculosis* H37Rv proteins, because IntPath(Zhou et al., 2012) also supports pathway analysis for this and other important pathogens. The pathway analysis on the *M. tuberculosis* H37Rv proteins involved in *H. sapiens*–*M. tuberculosis* H37Rv PPIs predicted by the stringent DDI-based approach is given in Table 4.6.

4.2.8 Analysis of domain properties of proteins involved in host–pathogen PPIs

The analysis of protein domain properties considers the number of domains and the degrees of domains on proteins. The protein domain properties directly reflect differences between the proteins involved in inter-species host–pathogen PPIN and intra-species PPIN. We analyze the domain properties of both *M. tuberculosis* H37Rv and *H. sapiens* involved in the predicted host–pathogen PPIs, and comparing them with other proteins in their own intra-species PPIN. As a control experiment, we also conduct the same analysis on the *H. sapiens* proteins in the gold standard *H. sapiens*–HIV PPIs(Fu et al., 2009) to see whether the *H. sapiens* proteins in the gold standard *H. sapiens*-HIV PPIs exhibit similar properties.

As the host–pathogen PPIs are predicted by the stringent DDI-based approach, to avoid biased analysis, we use a different domain annotation system in this analysis. The annotation of both *M. tuberculosis* H37Rv and *H. sapiens* protein domains is accomplished using HMMER-V3.0(Eddy, 2011). The domain profiles used in the protein domain annotation are Pfam-A(Bateman et al., 2004). The threshold for the domain annotation is E-value(iE-value) $\leq E - 20$ and accuracy ≥ 0.9. For each domain anno-

tated on each protein, we retrieve the sequences of these domains on every protein for the following analysis.

For the domain degree analysis, we obtain the DDI(Domain-Domain Interaction) data from the DOMINE database. DDIs "inferred from PDB entries" and "high confidence predictions" in the DOMINE database are considered in this study, while "medium confidence predictions" and "low confidence predictions" are discarded. For each domain, we count the number of interaction partners in the DOMINE database(only "inferred from PDB entries" and "high confidence predictions") as the degree of that domain. We analyze the above protein domain properties and summarize the results in Table 4.7.

4.2.9 Software Packages and Datasets

The software packages and database tools used in this study are:

- IntPath(Zhou et al., 2012)
- Cytoscape(Smoot et al., 2011)
- InterPro(Hunter et al., 2012)
- InterProScan(Quevillon et al., 2005)
- DAVID(Dennis Jr et al., 2003)

The datasets used in this study are:

- *M. tuberculosis* H37Rv PPI dataset consisting of four reliable subsets of the B2H PPI dataset and STRING PPI dataset(threshold at 770)(Zhou and Wong, 2011).
- *H. sapiens* physical PPI dataset collected from MINT(Zanzoni et al., 2002), BioGRID(Stark et al., 2011), and IntAct(Hermjakob et al., 2004); date of download is November 10, 2011.

- *S. cerevisiae* physical PPI dataset collected from MINT(Zanzoni et al., 2002), BioGRID(Stark et al., 2011), and IntAct(Hermjakob et al., 2004); date of download is November 10, 2011.
- Protein domain annotation (protein2ipr) from InterPro(Hunter et al., 2012); date of download is March 5th, 2012.
- DDI data from the 3did database(Stein et al., 2011)(version November 28, 2010).
- DDI data from the DOMINE database V2.0(Yellaboina et al., 2011).
- Pfam-A Domain profiles(Bateman et al., 2004).
- *H. sapiens*–HIV-1 PPI dataset downloaded from "HIV-1, human protein interaction database at NCBI"(Fu et al., 2009).

4.3 Results

4.3.1 Prediction of host–pathogen PPIs

Because of the stringent alignment threshold used for identifying query and template domain instances, lots of instances with large sequence variation under the same domain ID are filtered out, leaving very few domain instances for study. Also, our template interacting domain instances are from structurally resolved data in 3did, therefore the template domain instances are a relatively small number. Due to these two factors, our stringent DDI-based approach predicted PPI datasets are usually small. We have predicted 92 *H. sapiens-M. tuberculosis* H37Rv PPIs and this small set of predicted host–pathogen PPIs are analyzed using several approaches as discussed in the following sections. We visualize the predicted host–pathogen PPIN consisting of these 92 *H. sapiens-M. tuberculosis* H37Rv PPIs using Cytoscape(Smoot et al., 2011) in Figure 4.1. The orange dots are *M. tuberculosis* H37Rv proteins, while the blue dots are

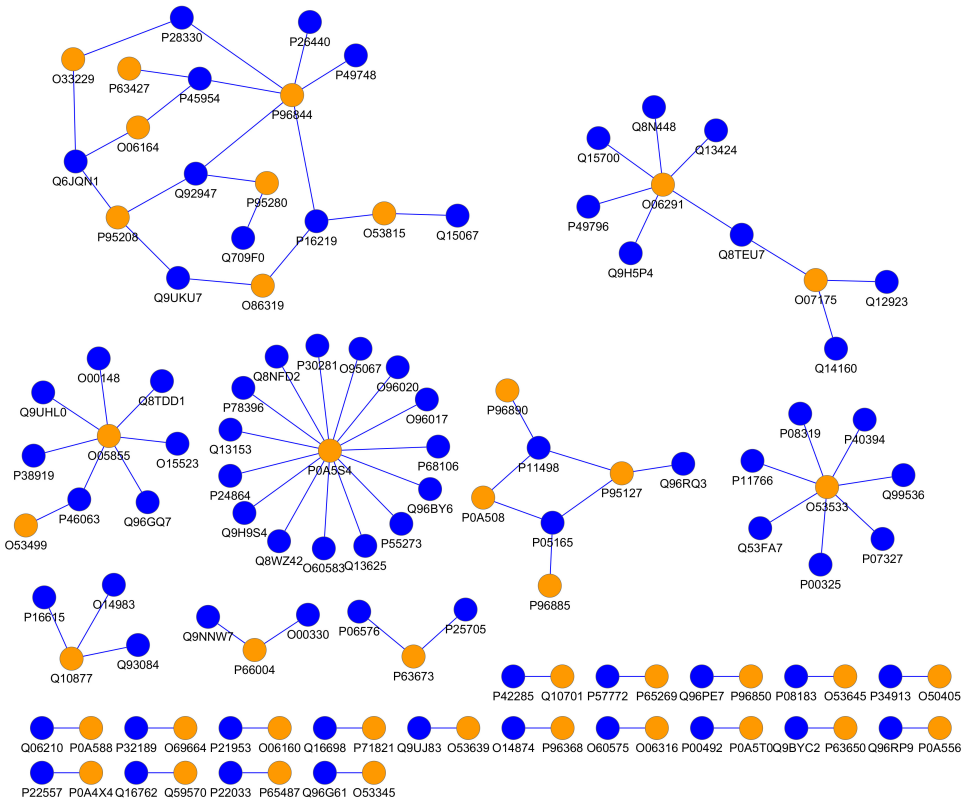


Figure 4.1: Visualization of predicted *H. sapiens–M. tuberculosis* H37Rv PPI network. The orange dots are *M. tuberculosis* H37Rv proteins, while the blue dots are *H. sapiens* proteins.

H. sapiens proteins. From Figure 4.1 we can observe that, like many host–pathogen PPINs, the pathogen proteins tend to be hubs in host–pathogen PPIN.

4.3.2 Prediction of intra-species PPIs

Currently no large-scale high-quality *H. sapiens-M. tuberculosis* H37Rv dataset is available. So we can not directly assess the performance of our stringent DDI-based approach in the inter-species host–pathogen system. Reluctantly, we turn to the intra-species system for the assessments. We predict intra-species *H. sapiens* PPIs using the stringent and the conventional DDI-based approaches. Altogether 839 *H. sapiens* PPIs

are predicted by the stringent DDI-based approach. In contrast, 724185 *H. sapiens* PPIs are predicted by the conventional DDI-based approach. Just from the number of PPIs predicted by two approaches the differences are obvious. Our stringent DDI-based approach relies on very high sequence similarity to the template domain instances and stands on the stringent domain instances to make the prediction. Therefore only a small amount of PPIs are predicted. And the small number of structurally resolved template interacting domain instances also limits the number of PPIs we can predict using our stringent DDI-based approach. Whereas the conventional DDI-based approach derives the possible interacting domain information from known PPI datasets(which can be abundant for some species), and treats all domain instances annotated under the same domain ID as the same. So a large number of PPIs can be predicted by the conventional DDI-based approach. We compare the performance of our stringent DDI-based approach and the conventional DDI-based approach based on gold standard PPI datasets and percentage of PPIs having coherent informative GO terms.

4.3.3 Assessment based on gold standard *H. sapiens* PPIs

We collect the known *H. sapiens* physical PPI datasets from MINT(Zanzoni et al., 2002), BioGRID(Stark et al., 2011), and IntAct(Hermjakob et al., 2004) as our gold standard PPI dataset to assess the *H. sapiens* PPIs predicted by the stringent and the conventional DDI-based approaches. We calculate and plot the precision-recall curve of the stringent and the conventional DDI-based approaches; see Figure 4.2. From the plots we can see both of the prediction approaches achieve better precision when the threshold increases. This shows that the scoring strategies adopted by both prediction approaches in calculating the "interaction strength" are valid in telling the likelihood of predicted PPIs being real. From the precision-recall curves, one can clearly tell that overall the stringent DDI-based approach consistently predicts PPIs with much higher precision than that of the conventional DDI-based approach.

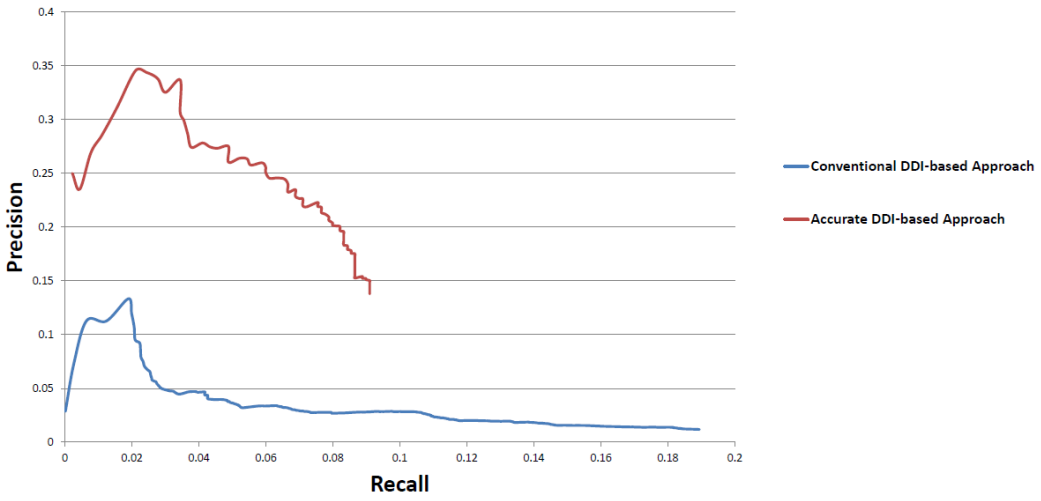


Figure 4.2: Assessment of the stringent and the conventional DDI-based approaches through gold standard *H. sapiens* PPIs. We plot the precision-recall curve.

As the conventional DDI-based approach makes a large number of predictions, it has higher recall. The precision-recall curve shows that our stringent DDI-based approach can only predict small amount of PPIs but with much higher accuracy than the conventional approach. As the two approaches predict very different number of PPIs, we also choose some special points to compare the performance of the two prediction approaches, see Table 4.1.

We can see that when our stringent DDI-based approach predicts 839 *H. sapiens* PPIs, 82 of which overlap with the gold standard; when the conventional DDI-based approach predicts 885 *H. sapiens* PPIs, only 11 of which overlap with the gold standard. The conventional DDI-based approach has to predict 3085 *H. sapiens* PPIs in order to have 81 *H. sapiens* PPIs overlapping with the gold standard. All these assessments using the gold standard *H. sapiens* PPIs clearly show that our stringent DDI-based approach is more accurate and has better performance than that of the conventional DDI-based approach.

Conventional DDI-based Approach	Overlap with Gold Standard
Top 3085 PPIs	81
Top 885 PPIs	11
Stringent DDI-based Approach	Overlap with Gold Standard
All 839 PPIs	82

Table 4.1: Assessment of the stringent and the conventional DDI-based approaches through gold standard *H. sapiens* PPIs. This table summarizes the assessment of the stringent and the conventional DDI-based approaches through gold standard human PPIs. In order for the conventional DDI-based approach to attain an amount of overlap with gold standard human PPIs similar to the stringent DDI-based approach, a much larger number of (false positive) predicted PPIs must be accepted. Conversely, if the conventional DDI-based approach is restricted to a similar number of predictions as the stringent DDI-based approach, a much lower overlap with gold standard human PPIs must be accepted.

4.3.4 Assessment based on coherent informative GO annotation of predicted *H. sapiens* PPIs

To further compare the performance of the stringent and the conventional DDI-based approaches, we calculate the percentage of PPIs that have coherent informative GO terms. From Figure 4.3 and Figure 4.4, the overall percentage of PPIs having coherent informative GO terms reveals that both approaches work well—as moving towards to a higher threshold (smaller number of top PPIs) leads to a higher percentage of PPIs having coherent informative GO terms.

As shown in Figure 4.3, the PPI dataset predicted by our stringent DDI-based approach starts with high percentage of PPIs having coherent informative GO terms; this indicates overall good performance as the PPI dataset predicted by our stringent DDI-based approach has low noise level and high quality. In contrast, the PPI dataset predicted by the conventional DDI-based approach does not show as good performance as the stringent DDI-based approach in terms of the overall percentage of PPIs having coherent informative GO terms—the PPI dataset predicted by the conventional DDI-based approach starts with a low percentage of PPIs having coherent informative GO terms, especially very low percentage of cellular compartment (CC) terms and biological

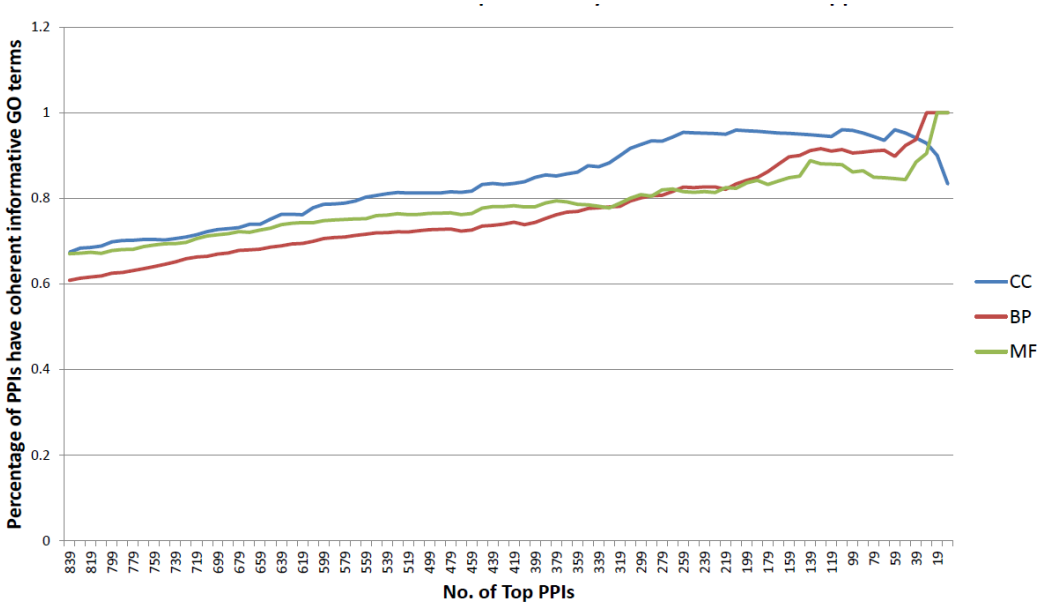


Figure 4.3: Informative GO assessment of the PPIs predicted by the stringent DDI-based approach. Informative GO assessment of the PPIs predicted by the stringent DDI-based approach.

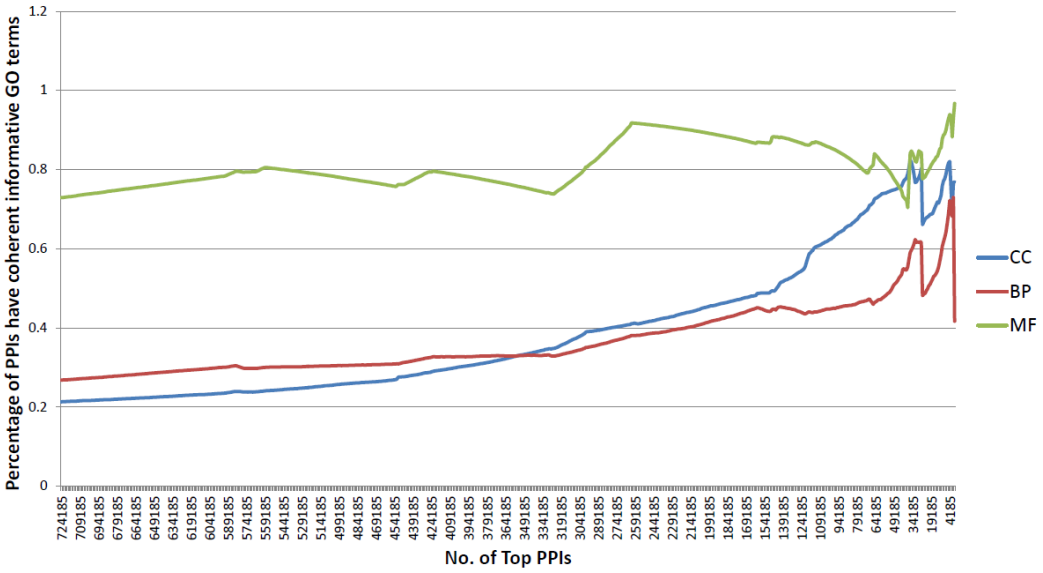


Figure 4.4: Informative GO assessment of the PPIs predicted by the conventional DDI-based approach. Informative GO assessment of the PPIs predicted by the conventional DDI-based approach.

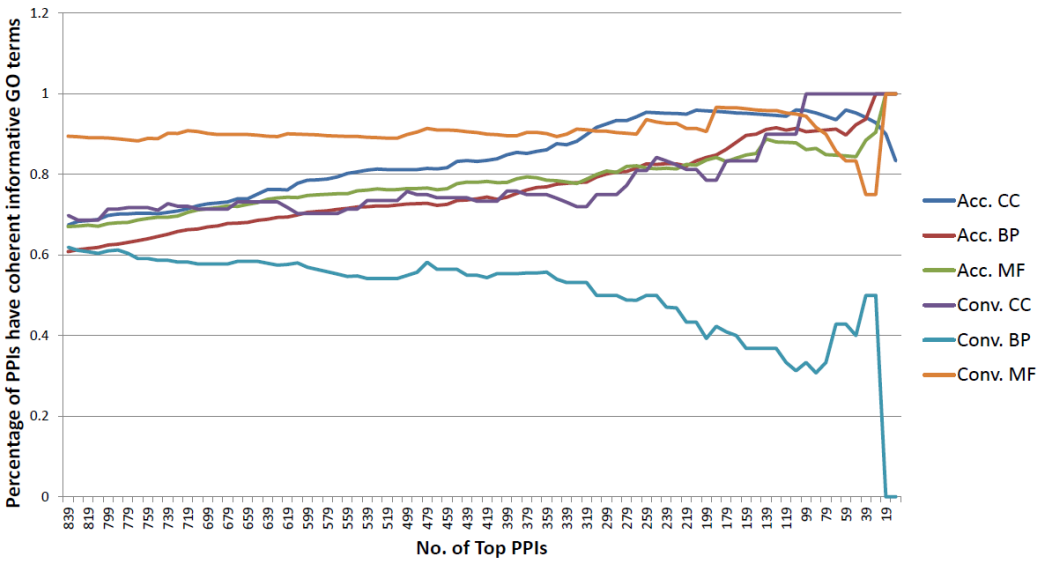


Figure 4.5: Informative GO assessment of the top 839 PPIs predicted by the stringent and the conventional DDI-based approaches. Informative GO assessment of the top 839 PPIs predicted by the stringent and the conventional DDI-based approaches. "Acc." means the PPIs predicted by the stringent DDI-based approach; "Conv." means the PPIs predicted by the conventional DDI-based approach.

process (BP) terms; this indicates that the PPI dataset predicted by the conventional DDI-based approach has high noise and the quality is not good.

As the PPI datasets predicted by the two approaches are very different in the number of predicted PPIs, it may not be a sufficient assessment seeing only overall plots of percentage of PPIs having coherent informative GO terms. Therefore, we focus on the top 839 PPIs respectively predicted by the stringent and the conventional DDI-based approaches and plot their percentage of PPIs having coherent informative GO terms in Figure 4.5.

We can clearly observe that PPIs predicted by the stringent DDI-based approach have consistently higher percentage of coherent informative CC and BP terms; see Figure 4.5.

The percentage of PPIs that have coherent informative GO terms may also be influenced by the number of GO terms that are annotated to the proteins in the PPI

Conventional DDI-based Approach	CC term No.	BP term No.	MF term No.
All 724185 PPIs	140	880	247
Top 839 PPIs	28	94	34
Stringent DDI-based Approach	CC term No.	BP term No.	MF term No.
All 839 PPIs	116	820	237

Table 4.2: Number of informative GO terms annotated to proteins involved in PPIs predicted by the stringent and the conventional DDI-based approach. This table summarizes the number of informative GO terms annotated to proteins involved in PPIs predicted by the stringent and the conventional DDI-based approach.

datasets. So we summarize the number of GO terms that are annotated to proteins in all 839 PPIs predicted by the stringent DDI-based approach, and proteins in all 724185 PPIs and the top 839 PPIs predicted by the conventional DDI-based approach in Table 4.2.

This table shows that although a high percentage of the PPIs predicted at a high threshold by the conventional DDI-based approach has coherent informative GO terms, this may be due the fact that these top 839 PPIs are annotated with very few distinct GO terms. Even with such a smaller number of informative GO terms we can see that the percentage of PPIs predicted by the conventional DDI-based approach having coherent informative GO terms is still consistently lower than the stringent DDI-based approach; this strongly supports the conclusion that the stringent DDI-based approach has a much better performance than that of the conventional DDI-based approach in predicting reliable PPIs.

4.3.5 Cellular compartment distribution of *H. sapiens* proteins targeted by predicted host–pathogen PPIs.

The cellular compartment distribution of the *H. sapiens* proteins targeted by the host–pathogen PPIs predicted by our stringent DDI-based approach is an important indicator of the performance of the prediction approach and the quality of the *H. sapiens-M. tuberculosis* H37Rv PPIs predicted. Host cellular compartments related to pathogen

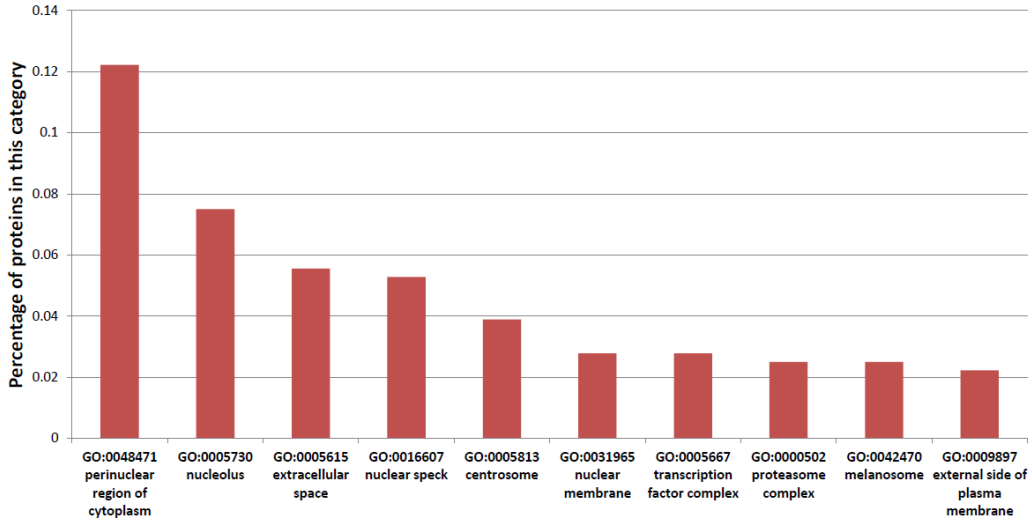


Figure 4.6: Cellular compartment distribution of *H. sapiens* proteins targeted by host–pathogen PPIs predicted by the stringent DDI-based approach. Cellular compartment distribution of *H. sapiens* proteins targeted by host–pathogen PPIs predicted by the stringent DDI-based approach.

infection that could be expected to be involved in PPIs with the pathogen, but not over-represented in the predicted set. Therefore, if the targeted *H. sapiens* proteins are mostly located in cellular compartments having a close relationship with pathogen infection then the predicted results are more convincing. We identify the informative CC terms in *H. sapiens* proteins. Then we calculate the number and percentage of proteins in the datasets that have been annotated with each of the informative CC terms. Then we plot the located informative CC terms for the targeted *H. sapiens* proteins by the stringent DDI-based approach in Figure 4.6, with detail statistics given in Table 4.3.

Many of the host–pathogen PPIs predicted by the stringent DDI-based approach target *H. sapiens* proteins are located in very relevant cellular compartments. *M. tuberculosis* H37Rv infection has a close relationship with mitochondria activities and function and induces quantitatively distinct changes in the mitochondrial proteome(Jamwal et al., 2013). Ultrastructural changes in the mitochondria and mitochondrial clus-

Cellular Compartment	Percentage(%)	No. of Proteins
GO:0005759 mitochondrial matrix	40.91%	18
GO:0005730 nucleolus	6.82%	3
GO:0045211 postsynaptic membrane	6.82%	3
GO:0005741 mitochondrial outer membrane	4.55%	2
GO:0016469 proton-transporting two-sector ATPase complex	4.55%	2
GO:0044439 peroxisomal part	4.55%	2
GO:0005813 centrosome	4.55%	2
GO:0031965 nuclear membrane	4.55%	2
GO:0048471 perinuclear region of cytoplasm	4.55%	2
GO:0019861 flagellum	2.27%	1
GO:0016324 apical plasma membrane	2.27%	1
GO:0005925 focal adhesion	2.27%	1
GO:0030027 lamellipodium	2.27%	1
GO:0035770 ribonucleoprotein granule	2.27%	1
GO:0016605 PML body	2.27%	1
GO:0016607 nuclear speck	2.27%	1
GO:0030018 Z disc	2.27%	1

Table 4.3: Cellular compartment distribution of *H. sapiens* proteins targeted by host–pathogen PPIs predicted by the stringent DDI-based approach. This table summarizes cellular compartment distribution of *H. sapiens* proteins targeted by host–pathogen PPIs predicted by the stringent DDI-based approach.

tering are also observed in the *M. tuberculosis* H37Rv infected cells(Jamwal et al., 2013). The augmentation of mitochondrial activity by *M. tuberculosis* H37Rv enables manipulation of host cellular mechanisms to inhibit apoptosis and ensure fortification against anti-microbial pathways(Jamwal et al., 2013). Therefore mitochondrial matrix(GO:0005759), mitochondrial outer membrane(GO:0005741) and proton-transporting two-sector ATPase complex(GO:0016469), are relevant to *M. tuberculosis* H37Rv infection.

H. sapiens proteins located at flagellum (GO:0019861) have much higher chance of interacting with *M. tuberculosis* H37Rv during infection as proteins located at flagellum are the first set of proteins that *M. tuberculosis* H37Rv comes across before invading the cell.

The CC term peroxisomal part(GO:0044439) is also strongly related to *M. tuberculosis* infection. It is found that the interaction between the mycobacterial phagosome and the endoplasmic reticulum leads to proteasome degradation and MHC class I pre-

sentation of *M. tuberculosis* antigens.

Focal adhesion(GO:0005925) is also closely interconnected to the *M. tuberculosis* infection process. In many bacterial pathogens, protein tyrosine phosphatases (PTPases) are essential for dephosphorylating host focal adhesion proteins and focal adhesion kinase. This dephosphorylation leads to destabilization of focal adhesions involved in the internalization of bacterial pathogens by eukaryotic cells(Persson et al., 1997; Black and Bliska, 1997). Therefore the proteins located at "Focal adhesion" compartment are very important target for *M. tuberculosis* infection of host. This strongly supports the validity of the prediction results of our stringent DDI-based approach.

The cellular compartment lamellipodium(GO:0030027) also supports the validity of our prediction results. It has been reported that host cell's actin filament network is interfered by pathogenic species of mycobateria(Guérin and de Chastellier, 2000b,a; Anes et al., 2003). A more recent study shows that *M. tuberculosis* affects actin polymerisation(Esposito et al., 2011).

The CC term nucleolus(GO:0005730) may also be related to *M. tuberculosis* infection, as *M. tuberculosis* infection of human macrophages blocks several responses to IFN-γ. The inhibitory effect of *M. tuberculosis* is directed at the transcription of IFN-γ-responsive genes(Ting et al., 1999). Several studies show that *M. tuberculosis* and its purified protein derivative induced HIV LTR primarily through transcriptional activation(Toossi et al., 1999).

The cellular compartment distribution analysis of the *H. sapiens* proteins targeted by host–pathogen PPIs strongly supports the validity of the PPI datasetpredicted by our stringent DDI-based approach.

GO terms	p-value
GO:0050660 FAD binding	2.27E-11
GO:0016462 pyrophosphatase activity	3.64E-06
GO:0004022 alcohol dehydrogenase (NAD) activity	8.70E-06
GO:0032559 adenyl ribonucleotide binding	9.27E-05
GO:0042626 ATPase activity, coupled to transmembrane movement of substances	6.54E-04
GO:0015405 P-P-bond-hydrolysis-driven transmembrane transporter activity	1.09E-03
GO:0042625 ATPase activity, coupled to transmembrane movement of ions	1.27E-03
GO:0000287 magnesium ion binding	8.04E-03
GO:0004466 long-chain-acyl-CoA dehydrogenase activity	1.28E-02
GO:0003960 NADPH:quinone reductase activity	2.55E-02
GO:0070402 NADPH binding	2.55E-02
GO:0004745 retinol dehydrogenase activity	6.25E-02
GO:0019841 retinol binding	7.45E-02
GO:0042288 MHC class I protein binding	9.81E-02

Table 4.4: Functional enrichment analysis of *H. sapiens* proteins involved in the host–pathogen PPI dataset predicted by the stringent DDI-based approach. This table summarizes the significantly enriched level 5 MF (Molecular Function) GO terms for *H. sapiens* proteins involved in the host–pathogen PPI dataset predicted by the stringent DDI-based approach. The analysis is produced using the DAVID database (threshold "count > 2, p-value < 0.1").

4.3.6 Functional enrichment analysis of proteins involved in host–pathogen PPIs

Functional enrichment analysis points out the possible functional relevance of *H. sapiens* proteins involved in the *H. sapiens-M. tuberculosis* H37Rv PPIN predicted by the stringent DDI-based approaches. The representative result—the most significantly enriched level 5 MF GO terms—is given in Table 4.4.

Most of the significantly enriched functional categories are strongly related to *M. tuberculosis* H37Rv infection, including adenyl ribonucleotide binding(GO:0032559), ATPase activity, coupled to transmembrane movement of substances (GO:0042626), P-P-bond-hydrolysis-driven transmembrane transporter activity(GO:0015405), ATPase activity, coupled to transmembrane movement of ions(GO:0042625), long-chain-acyl-CoA dehydrogenase activity(GO:0004466), NADPH:quinone reductase activity(GO:0003960), NADPH binding(GO:0070402), retinol dehydrogenase activity(GO:0004745), retinol binding(GO:0019841), and MHC class I protein binding(GO:0042288).

As described above, *M. tuberculosis* H37Rv infection is closely related to the mitochondria. Therefore all those MF terms closely related to mitochondria are relevant to *M. tuberculosis* H37Rv infection; the relevant GO terms include ATPase activity, coupled to transmembrane movement of substances (GO:0042626), P-P-bond-hydrolysis-driven transmembrane transporter activity(GO:0015405), ATPase activity, coupled to transmembrane movement of ions(GO:0042625), NADPH:quinone reductase activity(GO:0003960), NADPH binding(GO:0070402).

MHC class I protein binding(GO:0042288) is a strongly immune-related term which is also very relevant to *M. tuberculosis* H37Rv infection. Proteins enriched in this term play an important role in presenting *M. tuberculosis* antigens, which is essential for the immune response to this pathogen.

The long-chain-acyl-CoA dehydrogenase activity(GO:0004466) is a fatty acid-related term which is very relevant to *M. tuberculosis* H37Rv infection. Fatty acids and cholesterol appear to be the favored nutrients for *M. tuberculosis* inside *H. sapiens* cells(Lee et al., 2013). The breakdown of fatty acids and cholesterol can generate propionyl-CoA, which gives rise to potentially toxic intermediates(Lee et al., 2013). Through the methylcitrate cycle, the methylmalonyl pathway, or incorporation of the propionyl-CoA into methyl-branched lipids in the cell wall, *M. tuberculosis* expands the acetyl-CoA pool and alleviates the pressure from propionyl-CoA(Lee et al., 2013).

This functional enrichment analysis shows that our stringent DDI-based approach is accurate and has merits in identifying possible *H. sapiens* proteins that are involved in *H. sapiens–M. tuberculosis* H37Rv PPIs.

4.3.7 Pathway enrichment analysis of proteins involved in host–pathogen PPIs

Pathway enrichment analysis of the proteins involved in host–pathogen PPIN can provide rich information on the functional relevance of (both the host and pathogen)

Pathway names	p-value
Metabolic Pathways	4.82E-24
Fatty Acid Metabolism	4.04E-21
Valine, Leucine and Isoleucine Degradation	7.90E-19
Fatty Acid Beta Oxidation	5.00E-11
Glycolysis and Gluconeogenesis	4.84E-10
2-Oxobutanoate Degradation I	8.42E-10
p53 Signaling Pathway	3.86E-09
Ethanol Degradation II (cytosol)	5.92E-09

Table 4.5: Pathway enrichment analyses of *H. sapiens* proteins involved in the host–pathogen PPI dataset predicted by the stringent DDI-based approach. This Table shows the 8 most significantly enriched pathways for *H. sapiens* proteins involved in the host–pathogen PPI dataset predicted by our stringent DDI-based approach.

Pathway names	p-value
Fatty Acid β oxidation I	6.78E-3
Naphthalene degradation	7.29E-3

Table 4.6: Pathway enrichment analyses of *M. tuberculosis* H37Rv proteins involved in the host–pathogen PPI dataset predicted by the stringent DDI-based approach. This table summarizes the most significantly enriched pathways for *M. tuberculosis* H37Rv proteins involved in the host–pathogen PPI dataset predicted by our stringent DDI-based approach.

proteins involved in the host–pathogen PPIN. The analysis should show that the host proteins involved in host–pathogen interactions is a set of proteins that have functional correlation to pathways relevant to the pathogen's infection. Indeed *H. sapiens* proteins involved in the *H. sapiens-M. tuberculosis* H37Rv PPIN predicted by the stringent DDI-based approach are mostly enriched in the pathways are closely relevant to *M. tuberculosis* infection; see Table 4.5.

For example, "Fatty Acid Metabolism", "Fatty Acid Beta Oxidation", and "Glycolysis and Gluconeogenesis" are closely related to *M. tuberculosis* infection as fatty acids are one of the favored nutrients for *M. tuberculosis* inside *H. sapiens* cells(Lee et al., 2013). *M. tuberculosis* is able to grow on a variety of carbon sources, but mounting evidence has implicated fatty acids as the major source of carbon and energy for *M. tuberculosis* during infection(Marrero et al., 2010).

Also, *M. tuberculosis* switches its carbon source from sugars to fatty acids during the persistent phase of infection(Shi et al., 2010). Consequently, biosynthesis of sugars from intermediates of the tricarboxylic acid cycle is essential for its growth(Marrero et al., 2010). So the pathways "Metabolic Pathways", "Valine, Leucine and Isoleucine Degradation", "2-Oxobutanoate Degradation I", and "Ethanol Degradation II (cytosol)" maybe also be very related to *M. tuberculosis* infection as they are closely involved with intermediates of the tricarboxylic acid cycle which is essential for the growth of *M. tuberculosis*(Marrero et al., 2010). They may also contribute to the carbon flow of *M. tuberculosis* metabolism inside the human cell.

M. tuberculosis H37Rv proteins involved in the *H. sapiens-M. tuberculosis* H37Rv PPIN predicted by the stringent DDI-based approach are significantly enriched in the "Fatty Acid β oxidation I" pathway, see Table 4.6. This strongly supports the validity of our prediction results. As discussed above, fatty acids are the major source of carbon and energy for *M. tuberculosis* during infection(Marrero et al., 2010), and pathways involved with fatty acids metabolism strongly indicate association with the infection state of *M. tuberculosis* H37Rv. It is found that when the pathogen's acyl-coenzyme A synthetase gene is disrupted, infected mice survive significantly longer than those infected with the wild type, thus suggesting attenuation of the mutated pathogen. In fact the pathogen never attains the plateau phase of infection in mouse lungs when pathogen's acyl-coenzyme A synthetase gene is disrupted(Dunphy et al., 2010). *M. tuberculosis* fatty acyl-coenzyme A synthetase gene may serve to recycle mycolic acids for the long-term survival of the tubercle bacilli(Dunphy et al., 2010). Carbon rerouting is marked by a switch from metabolic pathways generating energy and biosynthetic precursors in growing bacilli to pathways for storage compound synthesis during growth arrest(Shi et al., 2010). This analysis result is in accord with the above cellular compartment distribution, functional enrichment analysis.

The presence of "Naphthalene Degradation" as a significant pathway is likely an ar-

tifact due to the sharing of many genes in this pathway with the "Fatty Acid β oxidation I" pathway. In particular, "naphthalene degradation" contains 40 genes, "Fatty Acid β oxidation I" contains 39 genes, and around half genes of these two pathways overlap with each other: 18 genes are mutually contained in both of these pathways. These shared genes are, Rv1934c, Rv0244c, Rv0972c, Rv3564, Rv3563, Rv3139, Rv2724c, Rv3274c, Rv3543c, Rv3505, Rv3504, Rv0215c, Rv0752c, Rv3560c, Rv0873, Rv1467c, Rv1933c, and Rv3544c. More specifically, there are 44 unique *M. tuberculosis* H37Rv proteins in the predicted *H. sapiens–M. tuberculosis* PPIs, 7 of these 44 proteins overlap with "Fatty Acid β oxidation I" (Rv2724c, Rv2500c, Rv3274c, Rv0752c, Rv0975c, Rv0400c, Rv3061c), 3 of these 44 proteins overlap with "naphthalene degradation" (Rv2724c, Rv3274c, Rv0752c), and all of these 3 proteins overlapping with "naphthalene degradation" are included in the 7 proteins overlapping with "Fatty Acid β oxidation I".

All the results support the validity of the *H. sapiens–M. tuberculosis* H37Rv PPIs predicted by our stringent DDI-based approach. Therefore the prediction results from our stringent DDI-based approach can serve as a reliable reference of PPIs between *H. sapiens* and *M. tuberculosis* H37Rv.

4.3.8 Analysis of domain properties of proteins involved in host–pathogen PPIs

We compare two domain properties of both *H. sapiens* and *M. tuberculosis* H37Rv proteins in the predicted *H. sapiens–M. tuberculosis* H37Rv PPIN and their own intra-species PPIN. We also conduct a similar analysis on *H. sapiens* proteins involved in the gold standard *H. sapiens*–HIV PPIN(Fu et al., 2009) as a control experiment. Table 4.7 provides summary results from the analysis of *H. sapiens* and *M. tuberculosis* H37Rv proteins.

It is obvious that *H. sapiens* proteins targeted by the predicted *H. sapiens–M. tuber-*

Organism	*H. sapiens* proteins		*H. sapiens* proteins	
PPIN	Hum-Mtb	Hum-Hum	Hum-HIV	Hum-Hum
Average No. of domains	1.79	1.31	1.42	1.27
P-value	4.40E-5		9.14E-17	
Average Domain degrees	17.95	10.22	13.23	9.21
P-value	1.79E-2		1.04E-10	

Table 4.7: Protein domain property analysis result. This table summarizes the protein domain analysis for *H. sapiens* proteins involved in the host–pathogen PPI dataset predicted by our stringent DDI-based approach comparing with the proteins involved in intra-species PPIN. Protein domain property analysis for *H. sapiens* proteins involved in gold standard *H. sapiens*–HIV PPI dataset(Fu et al., 2009) have also been conducted. In the table there are some abbreviations. Hum-Mtb: in predicted *H. sapiens–M. tuberculosis* H37Rv PPIN. Hum-Hum: in *H. sapiens* intra-species PPIN. Hum-HIV: in gold standard *H. sapiens*–HIV PPIN.

culosis H37Rv PPIN show properties very similar to those *H. sapiens* proteins targeted by the gold standard *H. sapiens*–HIV PPIN(Fu et al., 2009). This also supports the validity of our prediction results to some extent.

Both in the predicted *H. sapiens–M. tuberculosis* H37Rv PPIN and in the gold standard *H. sapiens*–HIV PPIN, *H. sapiens* proteins tend to have more domains and those domains tend to have higher degrees than those proteins in the intra-species *H. sapiens* PPIN.

The discoveries found by analyzing domain properties may be helpful in illuminating the basic mechanisms of how the host and pathogen proteins interact with each other, and may be useful in assessing the predicted host–pathogen PPIN.

4.4 Discussion

4.4.1 Sequence similarity between domain instances in DDI-based prediction

Comparing with conventional DDI-based approaches, our stringent DDI-based approach emphasizes the importance of domain instances in inferring interactions from

template DDIs. While this emphasis on stringent sequence similarity between template and query domain instances in transferring interaction results in significant improvement on prediction performance, it also draws attention to the large sequence variation among domain instances which may limit conventional DDI-based approaches.

4.4.2 Pros and cons of DDI-based prediction

The advantages of our stringent DDI-based approach have been discussed above, as it can predict more accurate PPIs on a small scale. The possible limitation of this approach is the lack of large-scale high-quality structurally-resolved DDIs. However, it is reasonable to expect more protein complex structures will be resolved, and the effectiveness of our stringent DDI-based approach will consequently be significantly strengthened.

Producing only a small amount of PPIs does not distract us from the merits of our stringent DDI-based approach, because the small number of highly accurate PPIs may already be more valuable than a huge amount of PPIs with a substantial fraction of noise. Highly accurate predicted PPIs, even though small in size, are usually very welcomed in experimental research, as they are a much more valuable reference for experimental verification than large datasets with high noise.

Accurate sequence alignment among domain instances are much more computationally expensive than the conventional DDI-based approach. This may limit the application of our stringent DDI-based approach to large-scale prediction of PPIs across many host–pathogen systems.

4.5 Conclusion

In this Chapter, we have proposed a stringent DDI-based prediction approach based on high sequence similarity between template domain instances and query domain instances. The assessment based on gold-standard *H. sapiens* PPIs and informative

GO annotation shows that the stringent DDI-based approach performs better than the conventional DDI-based approach. We have also predicted a small set of accurate *H. sapiens–M. tuberculosis* H37Rv PPIs. Through cellular compartment distribution, functional enrichment, and pathway enrichment analysis, we have demonstrated that this small set of accurate *H. sapiens–M. tuberculosis* H37Rv PPIs is valid and closely corresponds to *M. tuberculosis* H37Rv infection. This dataset of *H. sapiens–M. tuberculosis* H37Rv PPIs can be used for a variety of related studies as an important reference.

Chapter 5

Accurate Homology-Based Prediction

Homologs are a pair of proteins share a same ancestry. Homology-based prediction is one of the most frequently used computational approaches in predicting both intra-species and inter-species PPIs. However, some limitations are not properly resolved in several published works that predict eukaryote-prokaryote inter-species PPIs using intra-species template PPIs. We develop an accurate homology-based prediction approach by taking into account (i) differences between eukaryotic and prokaryotic proteins and (ii) differences between inter-species and intra-species PPI interfaces.

This accurate homology-based prediction approach provides an accurate strategy in predicting possible PPIs between eukaryotic hosts and prokaryotic pathogens. It performs better than a conventional homology-based approach in predicting PPIs between eukaryotic hosts and prokaryotic pathogens. The properties we have observed from the predicted *H. sapiens–M. tuberculosis* H37Rv PPI network are also important for understanding inter-species host–pathogen PPI networks and provide crucial novel insights for host–pathogen interaction studies.

5.1 Background

Homology-based approaches are the conventional way of predicting both intra-species and inter-species PPIs, with the assumption that the interaction between a pair of proteins in one species is likely to be conserved in related species(Matthews et al., 2001). They are also among the most frequently used methods in predicting host–pathogen PPIs, either being used alone(Lee et al., 2008; Krishnadev and Srinivasan, 2008; Tyagi et al., 2009; Krishnadev and Srinivasan, 2011) or in combination with other methods(Wuchty, 2011).

Current homology-based approaches generally transfer intra-species PPIs to predict host–pathogen PPIs. There are several limitations and concerns that have yet to be addressed. For example, (i) the protein-protein interaction interfaces between intra-species PPI and inter-species PPI are not exactly the same(Franzosa and Xia, 2011); (ii) the differences between prokaryotic and eukaryotic proteins are not considered. Therefore, the performance of conventional homology-based host–pathogen PPI prediction approaches is rather limited(Lee et al., 2008; Krishnadev and Srinivasan, 2008; Tyagi et al., 2009; Krishnadev and Srinivasan, 2011). In fact, most of these published works lack stringent verification, Thus, the accuracy of conventional homology-based approaches in predicting host–pathogen PPI is largely unknown.

In this Chapter, we develop a novel homology-based approach for predicting the *H. sapiens–M. tuberculosis* H37Rv PPIs by specifically transferring the eukaryote-prokaryote PPIs from an experimental human-bacteria template PPI dataset. Moreover, we adopt a more accurate method in identifying homologs between species by taking into account of genomic context. This prediction approach specifically addresses the limitations of the conventional homology-based approaches.

Cellular compartment distribution analysis, disease-related enrichment analysis, pathway enrichment analysis, and functional category enrichment analysis show that our predicted *H. sapiens–M. tuberculosis* H37Rv PPI dataset has good quality. These

analyses also demonstrate that our accurate homology-based approach have much better performance than a conventional homology-based approach. Therefore this accurate homology-based approach can be used for predicting host–pathogen PPIs in a variety of different eukaryote-prokaryote host–pathogen systems.

Based on primary sequence analysis and topological analysis of the predicted host–pathogen protein-protein interaction network (PPIN), we discover some interesting properties of both pathogen and host proteins participating in host–pathogen PPIs, including the tendency to be hubs in the intra-species PPIN, tendency to have smaller average shortest path length, tendency to be more hydrophilic, tendency to have longer sequences and more domains. Furthermore, the domains in the proteins involved in host–pathogen PPIN tend to have lower charge and tend to be more hydrophilic in comparison with other proteins in the intra-species PPIN.

5.2 Methods

Our accurate homology-based approach for predicting host–pathogen (*H. sapiens–M. tuberculosis* H37Rv) PPIs specifically transfers eukaryote–prokaryote (human–bacteria) PPIs from the PATRIC database(Gillespie et al., 2011). Cellular compartment distribution analysis, disease-related enrichment analysis, pathway enrichment analysis, and functional category enrichment analysis strongly support our prediction results and show that the predicted PPIs correspond to the *M. tuberculosis* H37Rv infection process.

In a control study, we use a conventional homology-based approach to predict possible host–pathogen (*H. sapiens–M. tuberculosis* H37Rv) PPIs. The same distribution and enrichment analyses are conducted on both results predicted by our accurate approach and the conventional approach.

The comparison shows that our accurate homology-based approach has better performance in predicting more relevant and meaningful host–pathogen PPI than the

conventional approach.

We further analyze some of the basic sequence properties of proteins involved in the host–pathogen PPIN comparing with the counterparts involved in intra-species PPIN by examining the sequences, domains, hydrophobicity scales, domain interaction degrees, electronic charge, etc. We also perform topological analysis to illuminate the intra-species topological properties of both the host and pathogen proteins involved in the predicted *H. sapiens–M. tuberculosis* H37Rv PPIN.

5.2.1 Prediction of host–pathogen PPI networks

Conventional homology-based approaches generally transfer intra-species PPIs to predict host–pathogen PPIs. That is, if a protein X in the host and a protein Y in the pathogen are respectively homologous to a pair of proteins X' and Y' which are known to interact in a third species, X and Y are predicted to interact.

In contrast, our accurate homology-based approach specifically transfers eukaryote-prokaryote inter-species PPIs to predict host–pathogen PPIs. Specifically, if a protein X in a eukaryotic host is known to interact with a protein Y' in a prokaryote species, and Y' is homologous to a protein Y in a prokaryotic pathogen, then we predict X and Y to interact. Moreover, to more accurately determine homologous proteins with conserved interactions, we use a homolog matching method that takes genomic context into consideration.

This accurate homology-based approach takes the followings into account: (i) the interface between intra- and inter-species PPI are not exactly the same(Franzosa and Xia, 2011); (ii) the differences between prokaryotic and eukaryotic proteins are also very obvious (post-transcriptional modifications, structures). Figure 5.1 shows differences between (a) a conventional homology-based prediction approach and (b) our approach.

For the accurate homology-based approach, we collect from the PATRIC database (Gillespie et al., 2011) the template eukaryote-prokaryote human-bacteria PPIs and the

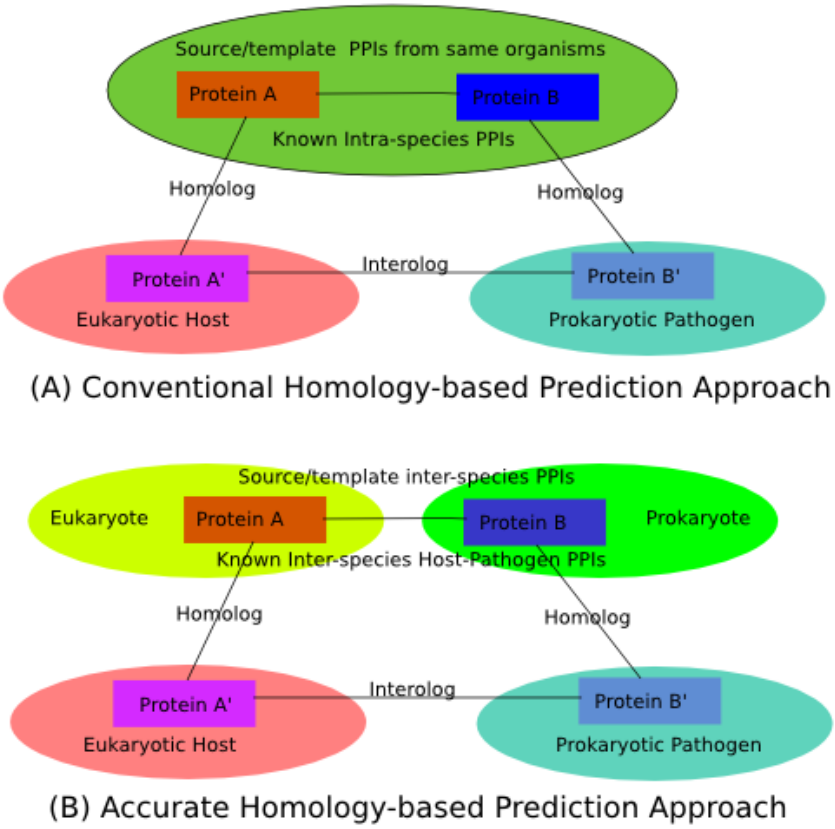


Figure 5.1: Representation of homology-based prediction approach. Representation of (A) the conventional homology-based prediction approach and (B)the accurate homology-based prediction approach adopted in this study.

genome sequences and gene feature files of relevant bacteria strains.

The list of bacteria strains in the PATRIC database (Gillespie et al., 2011) relevant to our study are *Bacillus anthracis str.* A2012, *Bacillus anthracis str.* Ames Ancestor, *Bacillus anthracis str.* Ames, *Bacillus anthracis str.* Sterne, *Francisella tularensis subsp tularensis* MA00-2987, *Francisella tularensis subsp tularensis* SCHU S4, *Shigella flexneri 2a str.* 301, *Yersinia pestis biovar Microtus* str. 91001, *Yersinia pestis* CO92, and *Yersinia pestis* KIM. These 10 major strains of bacteria cover 7120 PPIs in the PATRIC database, constituting 99% of the total PPIs contained in the database (data downloaded in April 3, 2012). The dataset collected above (PPIs between human and 10 major bacteria species) are the most abundant source eukaryote-prokaryote inter-species PPIs.

Our accurate homology-based prediction strategy works like this. If a human protein A is known to interact with a bacteria protein B in a template PPI (we call this template PPI a supporting template PPI), and the bacteria protein B has a homolog B' identified in *M.tuberculosis* H37Rv, then we predict that the human protein A and the *M.tuberculosis* H37Rv protein B' also interact with each other.

We count the number of supporting template PPIs as the "interaction strength" of each predicted *H. sapiens-M. tuberculosis* H37Rv PPI. This serves as one of the important parameters for evaluating how likely the predicted PPI is real compared with the rest of the predicted PPIs.

Using the accurate prediction approach as described above, we have predicted 1005 *H. sapiens-M. tuberculosis* H37Rv PPIs. We visualize the predicted network using Cytoscape(Smoot et al., 2011) in Figure 5.2.

We also predict host–pathogen PPIs using a conventional homology-based approach as a control experiment. Different from the accurate homology-based approach, the conventional homology-based approach uses template intra-species *H. sapiens* physical PPIs collected from three major PPI databases, MINT(Zanzoni et al., 2002), Bi-

oGRID(Stark et al., 2011), and IntAct(Hermjakob et al., 2004). All together 73251 *H. sapiens* physical PPIs are collected(data was downloaded on November 10, 2011). To predict *H. sapiens–M. tuberculosis* H37Rv PPIs using the conventional homology-based approach, we identify homologs between *H. sapiens* and *M. tuberculosis* H37Rv, and then transfer the intra-species *H. sapiens* PPIs to predict the inter-species *H. sapiens–M. tuberculosis* H37Rv PPIs.

The conventional homology-based prediction strategy just use different template PPIs for the prediction: if a human protein A interacts with a human protein B in a template PPI, and the human protein B has a homolog B' identified in *M.tuberculosis* H37Rv, then it predicts that the human protein A and the *M.tuberculosis* H37Rv protein B' interact with each other.

Using the conventional homology-based prediction approach as described above, we have predicted 326 *H. sapiens-M. tuberculosis* H37Rv PPIs.

To identify the homologs between *M.tuberculosis* H37Rv and the 10 bacteria (in our accurate approach) and also the between *M.tuberculosis* H37Rv and *H. sapiens* (in the conventional approach), we use the BBH-LS algorithm which computes positional homologs using both sequence and gene context similarity(Zhang and Leong, 2012). BBH-LS is considered to be a more accurate way of identifying homologs than other approaches which do not consider both the sequence and gene context similarity. The BBH-LS strength threshold β in this Chapter is set as 0.01.

5.2.2 Cellular compartment distribution of *H. sapiens* proteins targeted by the predicted host–pathogen PPIs.

The cellular compartment of the *H. sapiens* proteins targeted by the predicted host–pathogen PPIs are an important indicator of the quality of predicted PPIs. If the targeted *H. sapiens* proteins are located in cellular compartments that are very relevant to the pathogen's infection or are very likely to be involved in interactions with the

pathogen, then the result supports the host–pathogen predictions.

Gene Ontology (Cellular Compartment, CC) is one of the most comprehensive annotations for human proteins. Thus, we use it in our analysis. However, as the Gene Ontology is hierarchical, CC terms at the top levels may have more proteins annotated with them, while terms on lower levels may have less proteins annotate with them. Therefore, we only use the informative CC terms for our analysis. An informative CC term is defined here to be a term that has at least 90 proteins annotated to it, but each of its child terms has less than 90 proteins annotate with it. The cellular compartment distribution tells how many proteins(and the percentage) in the datasets that fall into each cellular compartment. We choose the top 10 frequently located cellular compartments of the *H. sapiens* proteins that are targeted by the accurate and the conventional homology-based prediction approaches. The results are shown in Table 5.1, 5.2 and Figure 5.3, 5.4.

5.2.3 Disease-related enrichment analysis of proteins involved in host–pathogen PPIs

Currently large-scale high-quality experimental *H. sapiens–M. tuberculosis* H37Rv PPIs are not readily available. Therefore a gold standard PPI dataset for assessing the predicted *H. sapiens–M. tuberculosis* H37Rv PPIs is not possible at the moment. However, there are several studies that examine *H. sapiens* gene expression profiles during *M. tuberculosis* H37Rv infection and treatment(Cliff et al., 2013; Chaussabel et al., 2003).

We obtain several *H. sapiens* gene lists related to *M. tuberculosis* H37Rv infection and treatment from the two studies(Cliff et al., 2013; Chaussabel et al., 2003). Chaussabel et al. (2003) identified the unique gene expression profiles of human macrophages and dendritic cells to phylogenetically distinct parasites, including *M. tuberculosis* H37Rv. We name this gene list "Macrophages and dendritic differentially expressed genes"; it contains 1531 differentially expressed *H. sapiens* genes. In another study,

Cliff et al. (2013) identifies several lists of blood gene expression profiles through tuberculosis treatment in different phases. Genes differentially expressed between diagnosis and week 1 of treatment are called "Early Changers", comprising 470 differentially expressed *H. sapiens* genes. Genes differentially expressed between week 4 and week 26 of treatment are called "Late Changers", comprising 327 differentially expressed *H. sapiens* genes. Genes which maintained a consistent pattern of change of gene expression and did not revert are called "Consistent Changers", comprising 406 differentially expressed *H. sapiens* genes.

Monocyte-derived dendritic cells and macrophages generated *in vitro* from the same individual blood donors were exposed to pathogens(*M. tuberculosis*), and gene expression profiles were assessed by microarray analysis in the work of Chaussabel et al. (2003). The genes differentially expressed during the exposure of pathogens are consistent with the concept that antigen-presenting cells have specific genes for use in the response to pathogens like *M. tuberculosis*(Chaussabel et al., 2003). Therefore the list of genes differentially expressed when the dendritic cells and macrophages are expose to *M. tuberculosis* are the enriched list of gene candidates that may have high possibility of involving in *H. sapiens–M. tuberculosis* H37Rv PPIs.

In the work of Cliff et al. (2013), *ex vivo* blood samples were collected from 27 first-episode pulmonary tuberculosis patients prior to starting standard therapy and after 1, 2, 4, and 26 weeks of successful treatment. Genome-wide gene expression profiles were obtained from *ex vivo* blood samples, the differentially expressed genes in different phases are called Early Changers, Late Changers and Constant Changers. The fast initial down-regulation of expression of inflammatory mediators coincided with rapid killing of actively dividing bacilli, whereas slower delayed changes occurred as drugs acted on dormant bacilli and coincided with lung pathology resolution(Cliff et al., 2013). As the drugs are working on killing the bacilli (*M. tuberculosis*), the differentially expressed genes at different phases correspond to the response to different

groups of *M. tuberculosis*(actively dividing bacilli, dormant bacilli, etc.).

These disease gene lists have also been used in assessments of predicted host–pathogen PPIs in other studies(Davis et al., 2007). These lists of differentially expressed genes form our reference disease-related gene lists. We conduct, against these disease-related gene lists, the enrichment(over-representation) analysis of the *H. sapiens* proteins involved in *H. sapiens–M. tuberculosis* H37Rv PPIs predicted by our accurate homology-based approach and by the conventional homology-based approach. The enrichment analysis uses hypergeometric test. The results are given in Table 5.3, Table 5.4.

5.2.4 Functional enrichment analysis of proteins involved in host–pathogen PPIs

Functional enrichment analysis is very important for identifying the functional relevance of the proteins involved in the host–pathogen PPIs. The presence of enriched(over-represented) functional categories that are closely related to pathogen infection, immune response, etc. serves as an important support for the validity of the prediction results.

The Gene Ontology (Molecular Function, MF) is one of the most comprehensive functional categories annotation. Therefore we conduct MF term enrichment analysis on the *H. sapiens* proteins involved in the predicted *H. sapiens-M. tuberculosis* H37Rv PPIs.

In this Chapter, we use the DAVID database(Dennis Jr et al., 2003) for the GO term enrichment analysis on the *H. sapiens* proteins involved in host–pathogen PPIs predicted by our accurate homology-based approach and the conventional homology-based approach. Representative results (significantly enriched level 5 MF terms, threshold "count $>$ 2, p-value $<$ 0.01") are shown in Table 5.5 and 5.6 (threshold "count $>$ 2, p-value $<$ 0.01").

DAVID does not support the functional enrichment analysis of *M. tuberculosis* H37Rv proteins. Moreover, as we have found in Chapter 5, most of the GO annotations for *M. tuberculosis* H37Rv are not specific enough to provide effective functional enrichment analysis. Therefore the functional analysis of *M. tuberculosis* H37Rv proteins is not discussed in this Chapter.

5.2.5 Pathway enrichment analysis of proteins involved in host–pathogen PPIs

Pathway data are a primary functional source for identifying a list of proteins' related functions. Usually for a set of proteins, if they are significantly enriched in certain pathways, it is very likely that this set of proteins play similar roles *in vivo*. Therefore pathway enrichment analysis is one of the most frequently used assessments on predicted host–pathogen PPIs.

For pathway enrichment analysis, we use the IntPath database(Zhou et al., 2012), which is currently one of the most comprehensive integrated pathway databases. The "Identify Pathways" function in IntPath can specifically identify the pathway enrichment of an input gene list. The "Identify Pathways" function in IntPath adopts the hypergeometric test to identify the input gene list's over-representation(enrichment) in the pathways. For each *H. sapiens* protein set (predicted by the accurate and the conventional homology-based approaches), we analyze the *H. sapiens* proteins' pathway enrichment using the IntPath database(Zhou et al., 2012), and the top 20 most significantly enriched pathways are listed in the Table 5.7 and 5.8. The enrichment analysis results summarized in the Table 5.7 and Table 5.8 provide an important evidence on which of the two approaches can predict more *H. sapiens–M. tuberculosis* H37Rv PPIs that are more relevant to *M. tuberculosis* H37Rv infection.

Besides comparing the quality of the two host–pathogen PPI datasets predicted by the two approaches based on pathway enrichment, we also analyze the pathway enrich-

ments for the *M. tuberculosis* H37Rv proteins. It is enabled by IntPath(Zhou et al., 2012), which supports pathway analysis for the important pathogens. The pathway analysis on the *M. tuberculosis* H37Rv proteins are not used to assess the performance of the two homology-based approaches—this is the first work to analyze the pathway enrichment of the pathogen proteins, so we have no base line to compare with. The results of pathway enrichment analysis on the *M. tuberculosis* H37Rv proteins involved in *H. sapiens–M. tuberculosis* H37Rv PPIs predicted by the accurate homology-based approach are listed in Table 5.9.

5.2.6 Analysis of sequence properties of proteins involved in host–pathogen PPIs

The analysis of primary protein sequence properties considers protein sequence length, number of domains, degrees of domains on proteins, length of domains on proteins, hydrophobicity, electron charge, etc. The protein sequence properties directly reflect differences between the proteins involved in inter-species host–pathogen PPIN and intra-species PPIN. We analyze the sequence properties of both *M. tuberculosis* H37Rv and *H. sapiens* involved in the predicted host–pathogen PPIs, and comparing them with other proteins in their own intra-species PPIN.

The annotation of both *M. tuberculosis* H37Rv and *H. sapiens* protein domains is accomplished using HMMER-V3.0(Eddy, 2011). The domain profiles used in the protein domain annotation are Pfam-A(Bateman et al., 2004). The threshold for the domain annotation is E-value(iE-value) $\leq E-20$ and accuracy ≥ 0.9. For each domain annotated on each protein, we retrieve the sequences of the domains on every protein for the following analyses.

Hydrophobicity of the proteins and domains are assessed based on the Kyte-Doolittle hydrophobicity scale. Kyte-Doolittle is a widely applied scale for delineating hydrophobic character of a protein. Regions with values above 0 are hydrophobic. We scan the

sequences of the proteins and domains and calculate the average hydrophobicity scale of each protein and each domain(sum the hydrophobicity scale of each amino acid and then divide the length of the protein/domain).

For the domain degree analysis, we obtain the DDI(Domain-Domain Interaction) data from the DOMINE database. DDIs "inferred from PDB entries" and "high confidence predictions" in the DOMINE database are considered in this study, while "medium confidence predictions" and "low confidence predictions" are discarded. For each domain, we count the number of interaction partners in the DOMINE database(only "inferred from PDB entries" and "high confidence predictions") as the degree of that domain.

The protein/domain net charge is calculated in the following ways: only three amino acids (Arginine, Histidine, Lysine) are positively charged (assigned value +1), two amino acids (Aspartic Acid, Glutamic Acid) are negatively charged (assigned value -1), the rest amino acid are neutral(assigned value 0). The average charge of each protein/domain is calculated by scanning the protein/domain sequence and take the average value of each protein/domain (sum the charge value divide the length of the protein/domain).

We analyze the above protein sequence properties and summarize the results in Table 5.10. We conduct a similar analysis on the domains, and the results are shown in Table 5.11.

5.2.7 Analysis of intra-species PPIN topological properties in host–pathogen PPIs

Intra-species PPIN topological properties was first examined and reported by Calderwood et al. (2007) and then repeatedly confirmed by others(Zhou et al., 2013). In this Chapter, we also conduct a similar study on the targeted *H. sapiens* proteins by examining the number of interaction partners in the intra-species PPIN. Previous analyses

are mainly constrained on the *H. sapiens* proteins as the *H. sapiens* PPIN is available, while most of the pathogen's intra-species PPIs are not available. Due to the work of Zhou and Wong (2011) on *M. tuberculosis* H37Rv intra-species PPIN, a high quality *M. tuberculosis* H37Rv PPI dataset is now available. Therefore this work is the first-ever study that examines the intra-species PPIN topological properties of the pathogen proteins involved in host–pathogen PPIs.

We mainly consider three important topological properties, Degree(the number of interaction partners in the intra-species PPIN), Betweenness Centrality(a measure of a node's centrality in a network, equal to the number of shortest paths from all vertices to all others that pass through that node in the intra-species PPIN), Shortest Path Length(average number of steps along the shortest paths for all possible pairs of network nodes, it measures the efficiency of information transport on a network). All these topological properties are calculated using Cytoscape's(Smoot et al., 2011) Analyze Network Plugin.

In this Chapter, *H. sapiens* intra-species PPIs are collected mainly from three databases, MINT(Zanzoni et al., 2002), BioGRID(Stark et al., 2011), and IntAct(Hermjakob et al., 2004). *M. tuberculosis* H37Rv PPIs are collected from STRING (with score above 770)(Szklarczyk et al., 2011) and the B2H PPI dataset(four small subsets of reliable PPIs)(Zhou and Wong, 2011). The results are shown in Table 5.12.

5.2.8 Software Packages and Datasets

The software packages and database tools used in this study are:

- IntPath(Zhou et al., 2012)
- BBH-LS(Zhang and Leong, 2012)
- Cytoscape(Smoot et al., 2011)
- HMMER-V3.0(Eddy, 2011)

- DAVID(Dennis Jr et al., 2003)

The datasets used in this study are:

- *M. tuberculosis* H37Rv PPI dataset consisting of four reliable subsets of the B2H PPI dataset and STRING PPI dataset(threshold at 770)(Zhou and Wong, 2011).
- *H. sapiens* PPI dataset collected from MINT(Zanzoni et al., 2002), BioGRID(Stark et al., 2011), and IntAct(Hermjakob et al., 2004), date of download is November 10, 2011.
- host–pathogen PPI data from PATRIC(Gillespie et al., 2011), date of download is April 3, 2012.
- 10 bacteria gene feature files, and whole genome fasta files are from PATRIC(Gillespie et al., 2011), date of download is April 3rd, 2012.
- DDI data from DOMINE database V2.0(Yellaboina et al., 2011).
- Pfam-A Domain profiles.(Bateman et al., 2004)
- *H. sapiens*–HIV-1 PPI dataset downloaded from "HIV-1, human protein interaction database at NCBI"(Fu et al., 2009).

5.3 Results

5.3.1 Prediction of host–pathogen PPI network

For our accurate homology-based approach, the most abundant template eukaryote-prokaryote inter-species PPIs are between human and 10 major bacteria species (7120 PPIs), therefore when predicting the *H. sapiens–M. tuberculosis* H37Rv PPIs we only need to identify the prokaryotic homologs between template and targeted species in

this situation. We identify, using BBH-LS (strength threshold $\beta \geq 0.01$), the homologs between *M.tuberculosis* H37Rv and the 10 bacteria from the PATRIC database. Here in this Chapter we use the "interaction strength"(the number of supporting template PPIs) as one of the parameters to evaluate how likely a predicted PPI is real, compared to the other predicted PPIs. For example, if there are 3 template human-bacteria PPIs transferring to the same *H. sapiens–M. tuberculosis* H37Rv PPI, then the PPI's interaction strength is "3". A total of 1005 *H. sapiens–M. tuberculosis* H37Rv PPIs are transferred from 7120 eukaryote-prokaryote (human-pathogen) PPIs. A visualization of the *H. sapiens-M. tuberculosis* H37Rv PPIN are shown in Figure 5.2. The blue dots are *M. tuberculosis* H37Rv proteins, while the orange dots are *H. sapiens* proteins. The "thickness" of an edge corresponds to the "interaction strength" of each predicted H. sapiens–M. tuberculosis H37Rv PPI. The predicted *H. sapiens–M. tuberculosis* H37Rv PPI dataset can be found in the additional file A.3.

For the conventional homology-based approach we obtain 73251 template PPIs from MINT, BioGRID and IntAct. We identify the homologs between human and *M.tuberculosis* to transfer PPIs in the prediction. Using BBH-LS (strength threshold $\beta \geq 0.01$), we identify 355 homologs between *M.tuberculosis* H37Rv and *H. sapiens*. Using these 355 homologs, we predict 326 *H. sapiens–M. tuberculosis* H37Rv PPIs from the 73251 eukaryote-eukaryote (human-human) intra-species PPIs.

The number of templates we start from and the number of predicted PPIs are surprisingly different between the accurate homology-based approach and the conventional homology-based approach. Using the same system and threshold in identifying homologs and then transferring a template PPI to predict a target host–pathogen PPI, in the accurate homology-based approach, 1005 inter-species PPIs are predicted from 7120 template PPIs; while in the conventional homology-based approach, only 326 inter-species PPIs are predicted from 73251 template PPIs. This result shows that our accurate homology-based approach are more efficient in using the template PPIs than

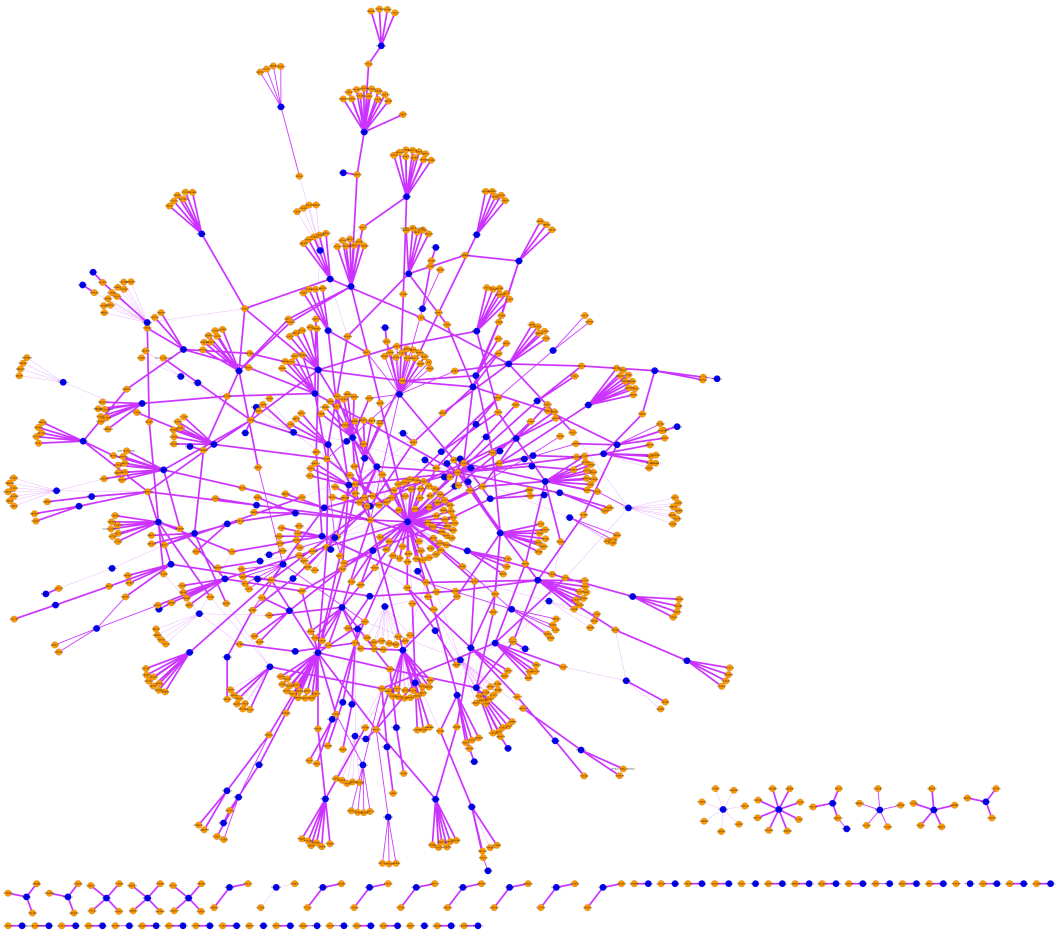


Figure 5.2: Visualization of the predicted *H. sapiens–M. tuberculosis* H37Rv PPI network. The blue dots are *M. tuberculosis* H37Rv proteins, while the orange dots are *H. sapiens* proteins. The "thickness" of an edge corresponds to the "interaction strength" of the predicted H. sapiens–M. tuberculosis H37Rv PPI, the thicker the edge the larger of the "interaction strength".

the conventional homology-based approach in predicting prokaryote-eukaryote inter-species PPIs. This highlights the merits of our accurate homology-based approach in applying to many host–pathogen systems.

5.3.2 Cellular compartment distribution of *H. sapiens* proteins targeted by predicted host–pathogen PPIs.

The cellular compartment of the *H. sapiens* proteins targeted by the predicted host–pathogen PPIs can provide important clues about the quality of the *H. sapiens-M. tuberculosis* H37Rv PPIs predicted. Host cellular compartments related to pathogen infection that could be expected to be involved in PPIs with the pathogen, but not over-represented in the predicted set. Therefore, if the targeted *H. sapiens* proteins are mostly located in cellular compartments having a close relationship with pathogen infection or known interactions with host cells, then the predicted results are solid.

We identify the informative CC terms of the *H. sapiens* proteins. Then we calculate the number and percentage of proteins in the datasets that have been annotated with each of the informative CC terms. Then we plot the top 10 most frequently located informative CC terms for the targeted *H. sapiens* proteins by the accurate and the conventional homology-based approach in Figure 5.3 and 5.4. We also summarize the statistics into Table 5.1 and 5.2.

Many of the host–pathogen PPIs predicted by the accurate homology-based approach target *H. sapiens* proteins locate in very relevant cellular compartments. This corresponds to the pathogen's infection and invasion of host cells. Among the top ten most frequent cellular compartment (GO) terms, four of them are closely relevant to the *M. tuberculosis* H37Rv infection. Those four terms are: extracellular space(GO:0005615), transcription factor complex(GO:0005667), proteasome complex(GO:0000502), external side of plasma membrane(GO:0009897).

H. sapiens proteins locate at extracellular space (GO:0005615) and extracellu-

Cellular Compartment	Percentage(%)	No. of Proteins
GO:0048471 perinuclear region of cytoplasm	12.2	44
GO:0005730 nucleolus	7.50	27
GO:0005615 extracellular space	5.56	20
GO:0016607 nuclear speck	5.28	19
GO:0005813 centrosome	3.89	14
GO:0031965 nuclear membrane	2.78	10
GO:0005667 transcription factor complex	2.78	10
GO:0000502 proteasome complex	2.50	9
GO:0042470 melanosome	2.50	9
GO:0009897 external side of plasma membrane	2.22	8

Table 5.1: Cellular compartment distribution of *H. sapiens* proteins targeted by the predicted host–pathogen PPIs. This table summarizes top 10 most frequent cellular compartments where the *H. sapiens* proteins(targeted by the accurate homology-based approach predicted host–pathogen PPIs) likely to be located in.

Cellular Compartment	Percentage(%)	No. of Proteins
GO:0048471 perinuclear region of cytoplasm	11.9	14
GO:0043025 neuronal cell body	5.93	7
GO:0005730 nucleolus	5.08	6
GO:0005759 mitochondrial matrix	5.08	6
GO:0016585 chromatin remodeling complex	4.24	5
GO:0005813 centrosome	3.39	4
GO:0005667 transcription factor complex	3.39	4
GO:0031965 nuclear membrane	3.39	4
GO:0017053 transcriptional repressor complex	2.54	3
GO:0005741 mitochondrial outer membrane	2.54	3

Table 5.2: Cellular compartment distribution of *H. sapiens* proteins targeted by the predicted host–pathogen PPIs. This table summarizes top 10 most frequent cellular compartments where the *H. sapiens* proteins(targeted by the conventional homology-based approach predicted host–pathogen PPIs) likely to be located in.

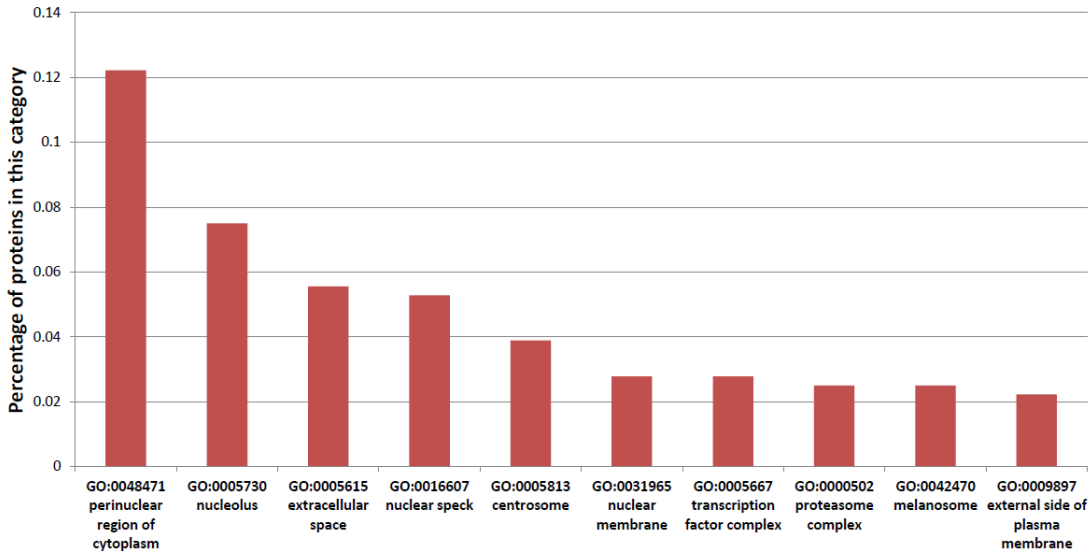


Figure 5.3: Cellular compartment distribution of *H. sapiens* proteins targeted by the accurate homology-based approach predicted host–pathogen PPIs. Cellular compartment distribution of *H. sapiens* proteins targeted by the accurate homology-based approach predicted host–pathogen PPIs(Top 10 cellular compartments).

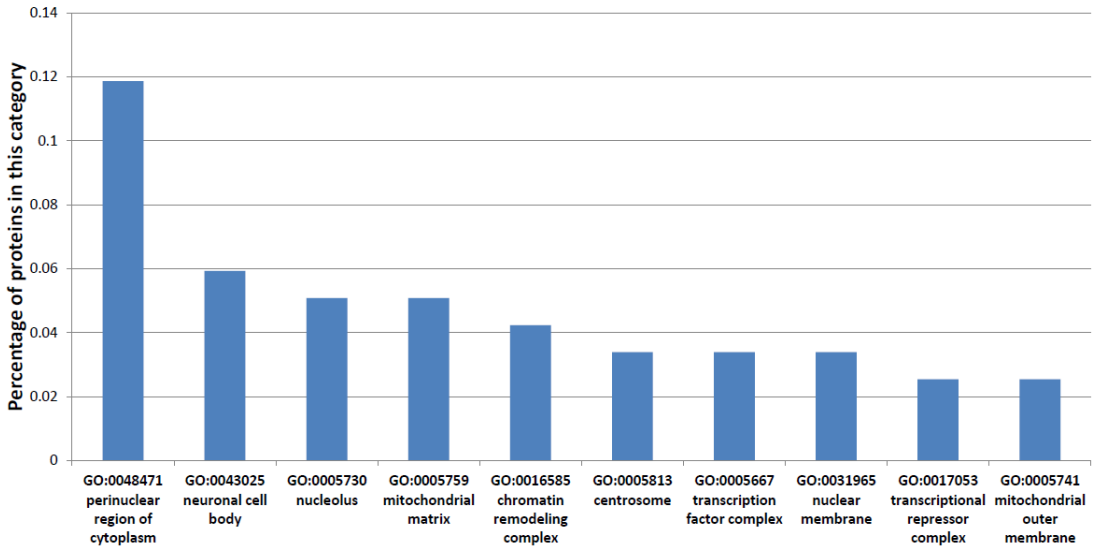


Figure 5.4: Cellular compartment distribution of *H. sapiens* proteins targeted by predicted host–pathogen PPIs(Top 10 Cellular Compartments).

lar space membrane (GO:0009897) have a much higher chance to interact with the pathogen *M. tuberculosis* H37Rv, because invasive bacteria pathogens are more likely to interact with the receptors, outer membrane proteins located on these two cellular compartments. The CC term, transcription factor complex(GO:0005667), is also relevant to *M. tuberculosis* infection, as *M. tuberculosis* has close interplay with *H. sapiens* cells on the transcription process.

For example, *M. tuberculosis* infection of human macrophages blocks several responses to IFN-γ. The inhibitory effect of *M. tuberculosis* is directed at the transcription of IFN-γ-responsive genes(Ting et al., 1999). There is a marked decrease in IFN-γ induced association of STAT1 with the transcriptional coactivators CREB-binding protein and p300 in *M. tuberculosis*-infected macrophages, indicating that *M. tuberculosis* directly or indirectly disrupts this protein-protein interaction that is essential for transcriptional responses to IFN-γ(Ting et al., 1999).

Several studies show that infection with *M. tuberculosis* increases the replication of HIV in mononuclear cells(Toossi et al., 1999). It turns out that *M. tuberculosis* and its purified protein derivative induced HIV LTR(Toossi et al., 1999). And the effect of *M. tuberculosis* and its purified protein derivative on HIV replication in monocytes is primarily one of transcriptional activation(Toossi et al., 1999). The CC term proteasome complex(GO:0000502), is also strongly related to *M. tuberculosis* infection. It is found that the interaction between the mycobacterial phagosome and the endoplasmic reticulum lead to proteasome degradation and MHC class I presentation of *M. tuberculosis* antigens. Thus, the results shown in Table 5.1 strongly supports the validity of our prediction results using the accurate homology-based prediction approach.

In contrast, there are three relevant CC terms out of the top ten most frequent cellular compartments where the conventional homology-based approach predicted host–pathogen PPIs targeted *H. sapiens* proteins locate at. These terms are: transcription factor complex (GO:0005667), mitochondrial matrix(GO:0005759), mitochondrial outer

membrane(GO:0005741); see Table 5.2.

M. tuberculosis H37Rv infection has a close relationship with mitochondria activities and function and was shown to induce quantitatively distinct changes in the mitochondrial proteome(Jamwal et al., 2013); therefore mitochondrial matrix(GO:0005759) and mitochondrial outer membrane(GO:0005741) are relevant to *M. tuberculosis* H37Rv infection. It is found that mitochondria in *M. tuberculosis* H37Rv-infected cells displayed robust activity with increased membrane potential and ATP synthesis(Jamwal et al., 2013). Ultrastructural changes in the mitochondria and mitochondrial clustering are also observed in the *M. tuberculosis* H37Rv infected cells(Jamwal et al., 2013). The augmentation of mitochondrial activity by *M. tuberculosis* H37Rv enables manipulation of host cellular mechanisms to inhibit apoptosis and ensure fortification against anti-microbial pathways(Jamwal et al., 2013).

From the results we can tell that the accurate homology-based approach has a better performance in predicting *H. sapiens-M. tuberculosis* H37Rv PPIs than the conventional homology-based approach.

5.3.3 Disease-related enrichment analysis of proteins involved in host–pathogen PPIs

The disease-related enrichment analysis results of *H. sapiens* proteins in *H. sapiens–M. tuberculosis* H37Rv PPIs predicted by the accurate homology-based approach show significant enrichment in all the gene lists, as summarized in Table 5.3. The significant enrichment of *H. sapiens* proteins involved in host–pathogen PPIs in "early, late, consistent changers" gene lists(Cliff et al., 2013) and also in "Macrophages and dendritic differentially expressed genes"(Chaussabel et al., 2003) is further evidence that *H. sapiens-M. tuberculosis* H37Rv PPIs predicted by our accurate homology-based approach are valid and very relevant to the infection process of *M. tuberculosis* H37Rv.

In contrast, the results from the conventional homology-based approach show much

Gene List	Overlap	p-value
Early Changers	32	1.022E-10
Late Changers	31	3.785E-14
Consistent Changers	35	1.500E-14
Early and Late Changers	56	6.996E-21
Early and Consistent Changers	49	3.721E-18
Consistent and Late Changers	42	1.499E-16
Macrophages and dendritic differentially expressed genes	107	2.097E-34

Table 5.3: Disease-related enrichment analysis of *H. sapiens* proteins involved in accurate homology-based approach predicted host–pathogen PPIs. This table summarizes *H. sapiens* proteins' (involved in the accurate homology-based approach predicted host–pathogen PPIs) enrichment (over-representation) in *M. tuberculosis* H37Rv infection and treatment-related differentially expressed gene lists.

less significant enrichment than the results from the accurate homology-based approach; see Table 5.4. This comparison clearly shows that our accurate homology-based approach has much better performance than the conventional homology-based approach.

The enrichment in disease-related genes observed above is perhaps not surprising. The template prokaryote-eukaryote PPIs from the source database used in this Chapter are mainly human-bacteria PPIs. Therefore, this may be a caveat. Unfortunately, due to the limited availability of non-pathogenic bacteria (i.e., commensals and probiotics) PPIs with human, we cannot overcome this problem by identifying a more neutral/unbiased source. This will be a good point to re-analyze when we have enough source templates from human-commensal and human-probiotic PPIs.

5.3.4 Functional enrichment analysis of proteins involved in host–pathogen PPIs

Functional enrichment analysis points out the possible functional relevance of *H. sapiens* proteins involved in the *H. sapiens-M. tuberculosis* H37Rv PPIN predicted by both the accurate and the conventional homology-based approaches. The representative result—the most significantly enriched level 5 MF GO terms—are listed in Table 5.5 and 5.6.

Gene List	Overlap	p-value
Early Changers	6	3.08E-02
Late Changers	6	6.11E-03
Consistent Changers	8	1.04E-03
Early and Late Changers	10	2.94E-03
Early and Consistent Changers	9	4.30E-03
Consistent and Late Changers	9	1.07E-03
Macrophages and dendritic differentially expressed genes	35	5.23E-14

Table 5.4: Disease-related enrichment analysis of *H. sapiens* proteins involved in conventional homology-based approach predicted host–pathogen PPIs. This table summarizes *H. sapiens* proteins' (involved in the conventional homology-based approach predicted host–pathogen PPIs) enrichment (over-representation) in *M. tuberculosis* H37Rv infection and treatment-related differentially expressed gene lists.

GO terms	p-value
GO:0051015 actin filament binding	6.12E-5
GO:0010843 promoter binding	5.76E-4
GO:0003713 transcription coactivator activity	7.18E-4
GO:0019901 protein kinase binding	3.63E-3
GO:0035257 nuclear hormone receptor binding	4.92E-3
GO:0070003 threonine-type peptidase activity	8.83E-3

Table 5.5: GO term enrichment analyses of *H. sapiens* proteins involved in the accurate homology-based approach predicted host–pathogen PPI dataset. It summarizes the most significantly enriched level 5 MF (Molecular Function) GO terms for *H. sapiens* proteins involved in the accurate homology-based approach predicted host–pathogen PPI dataset using DAVID database (threshold "count > 2, p-value < 0.01").

GO terms	p-value
GO:0003690 double-stranded DNA binding	8.11E-8
GO:0032559 adenyl ribonucleotide binding	1.54E-5
GO:0004672 protein kinase activity	2.50E-5
GO:0010843 promoter binding	1.08E-3
GO:0019901 protein kinase binding	4.13E-3
GO:0005031 tumor necrosis factor receptor activity	4.98E-3

Table 5.6: GO term enrichment analyses of *H. sapiens* proteins involved in the conventional homology-based approach predicted host–pathogen PPI dataset. It summarizes the most significantly enriched level 5 MF (Molecular Function) GO terms for *H. sapiens* proteins involved in the conventional homology-based approach predicted host–pathogen PPI dataset using DAVID database (threshold "count > 2, p-value < 0.01").

From the enrichment analysis results of the *H. sapiens* proteins targeted by the accurate homology-based approach predicted PPIs, shown in Table 5.5, five out of six significantly enriched terms are strongly *M. tuberculosis* H37Rv infection related functional categories, including "GO:0051015 actin filament binding", "GO:0010843 promoter binding", "GO:0003713 transcription coactivator activity", "GO:0019901 protein kinase binding", "GO:0035257 nuclear hormone receptor binding".

During vesicular fusion, the movement of endosomes and lysosomes are guided by the actin molecules associated with them. The fusion of endosomes with lysosomes is seriously affected by the disruption of actin filaments. And it has been reported that host cell's actin filament network are found to be interfered by pathogenic species of mycobateria(Guérin and de Chastellier, 2000b,a; Anes et al., 2003). A more recent study shows that *M. tuberculosis* affects actin polymerisation(Esposito et al., 2011). Therefore the functional enrichment analysis strongly supports the validity of the prediction results from our accurate homology-based approach, as the most enriched MF term shown in Table 5.5 is "actin filament binding"(GO:0051015). The significant enrichment of the terms "promoter binding(GO:0010843)","transcription coactivator activity(GO:0003713)" are closely related to *M. tuberculosis* infection, which also supports the validity of the prediction results by our accurate homology-based approach. As discussed above, *M. tuberculosis* infection of human macrophages has inhibitory effect on transcription of IFN-γ-responsive genes(Ting et al., 1999). It directly or indirectly influences transcriptional responses to IFN-γ(Ting et al., 1999). And *M. tuberculosis* increases the replication of HIV in mononuclear cells(Toossi et al., 1999). The effect of M. tuberculosis and its purified protein derivative on HIV replication in monocytes is primarily one of transcriptional activation(Toossi et al., 1999).

Bacterial pathogens have many ways to target one of the most ubiquitous signaling mechanisms of all eukaryotic host: phosphorylation by protein kinases(Krachler et al., 2011). MAPKs are evolutionarily conserved kinases that are important in cel-

lular signal transduction(Koul et al., 2004). There are three main families of MAPKs: (i)the c-Jun N-terminal kinases; (ii)the extracellular signal-related kinases; (iii)the p38 MAPK(Koul et al., 2004). Many bacterial pathogens (including *M. tuberculosis*) modify MAPK signalling to promote their survival in the host cells(Koul et al., 2004). By usurping p38 to interfere with CD1 surface expression, mycobacteria disrupt MAPK signaling pathways which play a crucial role in immune modulation(Gagliardi et al., 2009; Krachler et al., 2011). And p38 is exactly predicted to be targeted by *M. tuberculosis* H37Rv by our accurate homology-based approach. Therefore it is very reasonable and meaningful for the targeted host proteins to have significant functional enrichment in the term "GO:0019901 protein kinase binding". *M. tuberculosis* and its components are strong inducers of cytokines, such as tumour necrosis factor-alpha (TNF-α) and IL-1β(Valone et al., 1988; Wallis et al., 1986)

Many nuclear hormone receptors are shown to play a role in the repression of inflammatory mediators and they are also capable of modulating innate immunity in a positive manner(Chow et al., 2007). Liu et al. (2006) demonstrated, through the upregulation of VDR and vitamin D-1-hydroxylase genes, that TLRs adopt VDR antimicrobial activity in response to *M. tuberculosis* infection(Chow et al., 2007). Therefore the evidence is clear that, through positive and negative regulatory mechanisms, nuclear hormone receptors regulate innate immune responses to bacteria infections(Chow et al., 2007). This makes sense as this functional category of *H. sapiens* proteins are likely to be targeted by *M. tuberculosis* H37Rv proteins during infection.

In contrast, in the enrichment analysis results of *H. sapiens* proteins targeted by the conventional homology-based approach predicted PPIs, shown in Table 5.6, only four out of six significantly enriched terms are strongly *M. tuberculosis* H37Rv infection related functional categories, including "GO:0004672 protein kinase activity", "GO:0010843 promoter binding", "GO:0005031 tumor necrosis factor receptor activity", "GO:0019901 protein kinase binding".

This functional enrichment analysis shows that our accurate homology-based approach is accurate and has merits in identifying possible *H. sapiens* proteins that are involved in *H. sapiens–M. tuberculosis* H37Rv PPIs.

5.3.5 Pathway enrichment analysis of proteins involved in host–pathogen PPIs

Pathway enrichment analysis of the proteins involved in host–pathogen PPIN can tell a lot about the functional relevance of (both the host and pathogen) proteins involved in the host–pathogen PPIN. Therefore, pathway enrichment analysis has been used frequently in assessing host–pathogen PPI prediction results. The assessment stems from the basis that the host proteins involved in host–pathogen interactions should be a set of proteins that have functional correlation to pathways relevant to the pathogen's infection. So we also conduct pathway enrichment analysis to assess the quality of our prediction results and the performance of both the accurate and the conventional homology-based prediction approaches.

For *H. sapiens* proteins involved in the *H. sapiens-M. tuberculosis* H37Rv PPIN predicted by the accurate homology-based approach, they are mostly enriched in the pathways that are closely relevant to *M. tuberculosis* infection. Among the top 20 most significantly enriched pathways, 13 are closely relevant to *M. tuberculosis* infection; see Table 5.7. For example, "Amoebiasis", "Measles", "Tuberculosis", "Antigen processing and presentation", "Viral myocarditis", "Leishmaniasis", and "T cell receptor signaling pathway" are strongly infectious disease related and immune response related pathways which are obviously very relevant to *M. tuberculosis* infection. Moreover, our accurate homology-based approach predicted *H. sapiens* protein targets that are significantly enriched in the "Tuberculosis" pathway, which is a strong evidence supporting our prediction approach. "Focal adhesion", "Spliceosome", "Proteasome", "MAPK signaling pathway", and "Endocytosis" are essential pathways closely interconnected

to the "Tuberculosis" pathway. These essential pathways play crucial roles in the *M. tuberculosis* infection process and in the immune response to the infection.

The "Focal adhesion" pathway is closely interconnected to the *M. tuberculosis* infection process. In many bacterial pathogens, protein tyrosine phosphatases (PTPases) have been demonstrated to be essential for dephosphorylating host focal adhesion proteins and focal adhesion kinase. This dephosphorylation leads to destabilization of focal adhesions which are involved in the internalization of bacterial pathogens by eukaryotic cells(Persson et al., 1997; Black and Bliska, 1997). There are two functional PTPases in *M. tuberculosis*(Koul et al., 2000). A very interesting fact is that the *M. tuberculosis* genome lacks tyrosine kinases; so the existence of these two secretory tyrosine phosphatases (PTPases) shows that they are very likely involved in the dephosphorylation of host proteins. A study shows that, when the mptpB gene is deleted from *M. tuberculosis*, the mutant strain is attenuated in the lung and spleen of infected animals(Singh et al., 2003). Therefore the "Focal adhesion" pathway is a very important target for *M. tuberculosis* infection of host. The significant enrichment of this pathway strongly supports the validity of the prediction results of our accurate homology-based approach, as shown in Table 5.7.

The invasion of *M. tuberculosis* to the host cell is closely facilitated by endocytosis, which is one of early steps for the pathogen to interact with proteins inside the host cell.

Proteasome is also strongly related to *M. tuberculosis* infection. It is found that the interaction between the mycobacterial phagosome and the endoplasmic reticulum leads to proteasome degradation and MHC class I presentation of *M. tuberculosis* antigens.

MAPKs are evolutionarily conserved kinases that are important in cellular signal transduction(Koul et al., 2004). Many bacterial pathogens (including *M. tuberculosis*) modify MAPK signalling to promote their survival in the host cells(Koul et al., 2004).

From the biological aspect, the *H. sapiens* proteins involved in the *H. sapiens*-

M. tuberculosis H37Rv PPIs(predicted by the accurate homology-based approach) are highly likely to be involved in the above enriched pathways. This pathway enrichment analysis suggests that our accurate homology-based prediction accurately identifies *H. sapiens* proteins that are likely to be targeted by *M. tuberculosis* H37Rv.

In contrast, the pathway enrichment analysis of *H. sapiens* proteins involved in the *H. sapiens-M. tuberculosis* H37Rv PPIN predicted by the conventional homology-based approach shows that the conventional homology-based approach does not have the same good performance as the accurate homology-based approach. Among the top 20 most significantly enriched pathways, only 9 are closely relevant to *M. tuberculosis* infection; see Table 4(b). For example, "Hepatitis C", "Shigellosis", "T cell receptor signaling pathway", "EBV LMP1 signaling", and "Chagas disease (American trypanosomiasis)" are infectious disease related and immune response related pathways relevant to *M. tuberculosis* infection. "Endocytosis", "MAPK signaling pathway", "Apoptosis", and "Proteasome" are essential pathways also considered as related pathways.

This comparative analysis shows both homology-based approaches can predict the *H. sapiens-M. tuberculosis* H37Rv PPIN and pathway enrichment analysis supports both prediction results. However, the accurate homology-based approach has better performance than the conventional homology-based approach.

Among the most significantly enriched pathways, our accurate homology-based approach recovers the "Tuberculosis" pathway. We use the KEGG pathway map(Ogata et al., 1999) to visualize the *H. sapiens* proteins that have been targeted in our prediction results(in pink color) and all rest of the proteins participating in the pathway(in green color). The pathway map is shown in Figure 5.5.

Some cancer-related pathways are also present in the list of most enriched pathways. The presence of cancer pathways may or may not be regarded as artifacts of the pathway analysis. On one hand, cancers share lots of similarity with pathogen infection, including evading immune response, inducing apoptosis, metastasis and invading

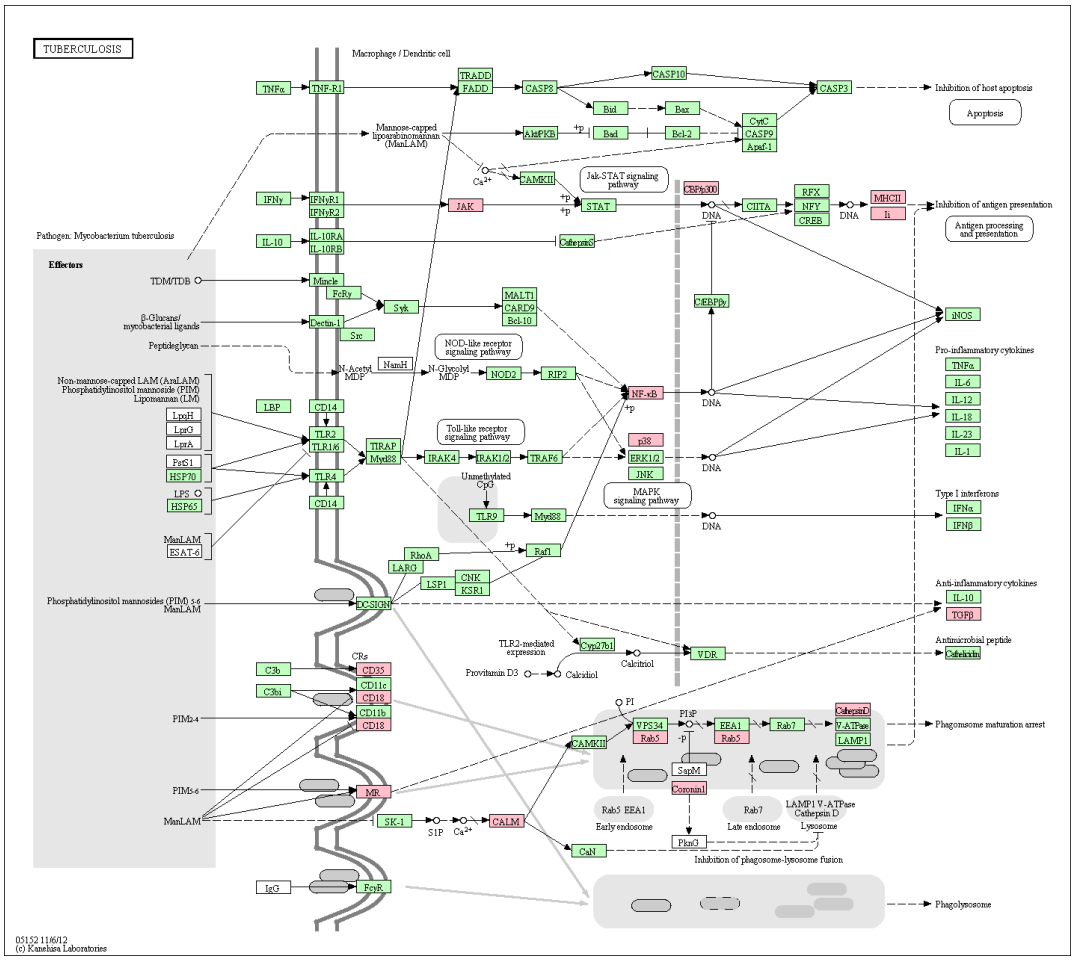


Figure 5.5: Visualization of the KEGG "Tuberculosis" pathway with *H. sapiens* proteins recovered by our predicted *H. sapiens–M. tuberculosis* H37Rv PPI network. The pink squares are *H. sapiens* proteins targeted in our predicted *H. sapiens–M. tuberculosis* H37Rv PPIN that are in the KEGG "Tuberculosis" pathway map. The green squares are *H. sapiens* proteins in the "Tuberculosis" pathway, but not recovered in our prediction.

Pathway names	p-value
Focal adhesion	5.85E-13
Translation factors	6.61E-12
Pathways in cancer	7.51E-12
Measles	5.21E-09
Pancreatic cancer	7.44E-09
Proteasome	8.80E-09
Antigen processing and presentation	1.68E-08
Adipogenesis	3.41E-08
Myometrial relaxation and contraction pathways	5.66E-08
MAPK signaling pathway	5.82E-08
Endocytosis	5.87E-08
Integrated cancer pathway	5.89E-08
Viral myocarditis	8.03E-08
Cell cycle	8.28E-08
Leishmaniasis	1.08E-07
T cell receptor signaling pathway	1.12E-07
Tuberculosis	2.76E-07
Spliceosome	7.79E-07
Renal cell carcinoma	7.82E-07
Amoebiasis	8.28E-07

Table 5.7: Pathway enrichment analysis of *H. sapiens* proteins involved in the accurate homology-based approach predicted host–pathogen PPI dataset. It summarizes the 20 most significantly enriched pathways for *H. sapiens* proteins involved in the host–pathogen PPI dataset predicted by our accurate homology-based approach.

the cells, etc. Therefore, many essential pathways that are highly interconnected to *M. tuberculosis* infection are also closely related to cancer pathways. Those essential pathways are "Apoptosis", "MAPK signaling pathway", "Jak-STAT signaling pathway", "Focal adhesion", etc.

On the other hand, the presence of cancer pathways in the set of highly enriched pathways is also caused by the overlap of many "core" proteins, which mostly are the housekeeping genes of *H. sapiens* cells.

M. tuberculosis H37Rv proteins involved in the accurate homology-based approach predicted *H. sapiens–M. tuberculosis* H37Rv PPIN are mostly enriched in pathways that are related to "general metabolism", "amino acid metabolism", "ribonucleotides metabolism", etc.; see Table 5.9. This makes sense as the pathogen infecting the human host undergoes rigorous metabolism in order to multiply and further infects the host.

Pathway names	p-value
Hepatitis C	2.03E-14
Pathways in cancer	2.52E-13
Endocytosis	3.20E-13
MAPK signaling pathway	5.66E-13
Neurotrophin signaling pathway	4.67E-12
Cell cycle	1.78E-11
Shigellosis	4.18E-11
T cell receptor signaling pathway	3.21E-10
Senescence and autophagy	7.20E-10
NOD-like receptor signaling pathway	9.06E-10
Prostate cancer	1.35E-09
EBV LMP1 signaling	4.64E-09
RIG-I-like receptor signaling pathway	4.74E-09
Acute myeloid leukemia	2.42E-08
Osteoclast differentiation	3.37E-08
Apoptosis	3.86E-08
Chagas disease (American trypanosomiasis)	9.86E-08
Pancreatic cancer	1.03E-07
Proteasome	1.14E-07
DNA damage response	1.25E-07

Table 5.8: Pathway enrichment analysis of *H. sapiens* proteins involved in the conventional homology-based approach predicted host–pathogen PPI dataset. It summarizes the 20 most significantly enriched pathways for *H. sapiens* proteins involved in the host–pathogen PPI dataset predicted by our conventional homology-based approach.

Pathway names	p-value
Metabolic pathways	6.81E-39
tRNA charging pathway	1.46E-18
Biosynthesis of secondary metabolites	1.54E-17
Pyrimidine metabolism	6.72E-10
Purine metabolism	2.25E-09
Aminoacyl-tRNA biosynthesis	6.47E-09
Alanine, aspartate and glutamate metabolism	3.09E-07
Superpathway of histidine, purine, and pyrimidine biosynthesis	3.25E-07
Superpathway of chorismate	1.14E-06
Arginine biosynthesis	1.39E-06
Superpathway of citrulline metabolism	2.13E-06
Tetrapyrrole biosynthesis I	2.13E-06
Tryptophan biosynthesis	2.13E-06
Phenylalanine, tyrosine and tryptophan biosynthesis	2.22E-06
Superpathway of cytosolic glycolysis, pyruvate dehydrogenase and TCA cycle	1.72E-05
Glyceraldehyde 3-phosphate degradation	3.47E-05
Gluconeogenesis I	3.92E-05
Pyrimidine ribonucleotides de novo biosynthesis	3.92E-05
Nucleotide excision repair	3.98E-05
Glycine, serine and threonine metabolism	4.53E-05

Table 5.9: Pathway enrichment analysis of *M. tuberculosis* H37Rv proteins involved in the predicted host–pathogen PPI dataset. This table summarizes the 15 most significantly enriched pathways for *M. tuberculosis* H37Rv proteins involved in the predicted host–pathogen PPI dataset.

Therefore the prediction results from our accurate homology-based approach can serve as a reliable reference of PPIs between *H. sapiens* and *M. tuberculosis* H37Rv.

This analysis result is in accord with the above cellular compartment distribution, disease gene list, pathway enrichment and functional category enrichment analysis results. All the results support the validity of the *H. sapiens–M. tuberculosis* H37Rv PPIs predicted by our accurate homology-based approach. Furthermore, all the analysis results above suggest our accurate homology-based approach has better performance than the conventional homology-based approach in predicting host–pathogen PPIs.

5.3.6 Analysis of protein sequence properties of proteins involved in host–pathogen PPIs

The analysis of the sequence properties of proteins involved in host–pathogen PPI network reveals many interesting properties that have not been reported before. In the analysis we compare several important features of both *H. sapiens* and *M. tuberculosis* H37Rv proteins/domains in the predicted *H. sapiens–M. tuberculosis* H37Rv PPIN and their own intra-species PPIN. Table 5.10 provides summary results from the analysis of *H. sapiens* and *M. tuberculosis* H37Rv proteins.

It is very obvious that in the predicted *H. sapiens–M. tuberculosis* H37Rv PPIN, *H. sapiens* proteins tend to have longer primary sequence, tend to have more domains, tend to have lower charge and tend to be more hydrophilic than those proteins in the intra-species *H. sapiens* PPIN. For *M. tuberculosis* H37Rv proteins, similar properties are also exhibited; for example, *M. tuberculosis* H37Rv proteins in the predicted *H. sapiens–M. tuberculosis* H37Rv PPIN tend to have longer primary sequences, tend to have more domains, tend to have lower charge and tend to be more hydrophilic than those proteins in the intra-species *M. tuberculosis* H37Rv PPIN. When we zoom in from the protein level to the domain level, we find the domains also exhibit similar properties as the proteins; see Table 5.11. The most significant properties for the domains in inter-

Organism	*H. sapiens* proteins		*M. tuberculosis* proteins	
PPIN	Hum-Mtb	Hum-Hum	Hum-Mtb	Mtb-Mtb
Average Length	769.3	623.0	486.0	328.7
P-value	1.33E-7		7.36E-17	
Average Hydrophobicity	-0.453	-0.413	-0.034	-0.027
P-value	2.39E-3		0.700	
Average Charge	0.058	0.065	0.068	0.079
P-value	9.07E-4		7.31E-7	
Average No. of domains	1.39	1.31	1.55	1.25
P-value	2.65E-2		2.82E-6	
Average Domain degrees	10.56	10.19	5.54	3.16
P-value	0.756		5.94E-4	

Table 5.10: Protein sequence properties analysis result. This table summarizes our analysis of protein sequence properties for *H. sapiens* and *M. tuberculosis* H37Rv proteins involved in the predicted host–pathogen PPI dataset compared with proteins involved in intra-species PPIN. In the table there are some abbreviations. Hum-Mtb: in predicted *H. sapiens–M. tuberculosis* H37Rv PPIN. Hum-Hum: in *H. sapiens* intra-species PPIN. Mtb-Mtb: in *M. tuberculosis* intra-species PPIN.

Organism	*H. sapiens* proteins		*M. tuberculosis* proteins	
PPIN	Hum-Mtb	Hum-Hum	Hum-Mtb	Mtb-Mtb
Average Length	205.0	188.4	210.0	187.2
P-value	0.863		2.04E-2	
Average Hydrophobicity	-0.355	-0.293	-0.033	0.037
P-value	2.15E-2		7.90E-4	
Average Charge	0.055	0.059	0.069	0.076
P-value	4.19E-2		9.93E-3	
Average degrees	11.66	11.62	4.42	4.47
P-value	0.97		0.89	

Table 5.11: Domain sequence properties analysis result. This table summarizes our analysis of domain sequence properties for *H. sapiens* and *M. tuberculosis* H37Rv proteins involved in the predicted host–pathogen PPI dataset, compared with proteins involved in intra-species PPIN. In the table there are some abbreviations. Hum-Mtb: in predicted *H. sapiens–M. tuberculosis* H37Rv PPIN. Hum-Hum: in *H. sapiens* intra-species PPIN. Mtb-Mtb: in *M. tuberculosis* intra-species PPIN.

Organism	*H. sapiens* proteins		*M. tuberculosis* proteins	
PPIN	Hum-Mtb	Hum-Hum	Hum-Mtb	Mtb-Mtb
Average Degree	26.69	12.56	25.67	16.16
P-value	2.18E-11		7.34E-9	
Average Betweeness Centrality	6.33E-4	8.23E-4	8.36E-3	1.63E-2
P-value	0.439		0.132	
Average Shortest Path Length	3.33	3.57	4.73	4.77
P-value	1.33E-30		0.65	

Table 5.12: Topological properties analysis result. This table summarizes our analysis of intra-species PPIN topological properties for *H. sapiens* and *M. tuberculosis* H37Rv proteins involved in the predicted host–pathogen PPI dataset, compared with proteins involved in intra-species PPIN. In the table there are some abbreviations. Hum-Mtb: in predicted *H. sapiens–M. tuberculosis* H37Rv PPIN. Hum-Hum: in *H. sapiens* intra-species PPIN. Mtb-Mtb: in *M. tuberculosis* intra-species PPIN.

species host–pathogen PPIN are that they tend to be more hydrophilic and tend to have lower charge than counterparts in the intra-species PPIN (both in *H. sapiens* and *M. tuberculosis* H37Rv proteins).

The discoveries found by analyzing sequence properties may be helpful in illuminating the basic mechanisms of how the host and pathogen proteins interact with each other, and may be useful in assessing the predicted host–pathogen PPIN.

5.3.7 Analysis of intra-species PPIN topological properties in host–pathogen PPIs

The results from the analysis of intra-species PPIN topological properties for *H. sapiens* and *M. tuberculosis* H37Rv proteins involved in the predicted host–pathogen PPI dataset in comparison with proteins involved in intra-species PPIN are summarized in Table 5.12.

From the intra-species PPIN topological properties of *H. sapiens* proteins involved in the predicted and gold standard host–pathogen PPINs, we conclude that *H. sapiens* proteins being targeted by pathogen proteins in the host–pathogen PPIs tend to have much higher degree than proteins in the intra-species PPIN. In other words, the host

proteins being targeted by pathogens are more likely to be hubs in their own intra-species PPIN. This result further strengthens the discoveries first reported by Calderwood et al. (2007) and is also in agreement with many studies that followed(Zhou et al., 2013).

In this Chapter we are the first to examine the intra-species PPIN topological properties of *M. tuberculosis* H37Rv proteins involved in the *H. sapiens–M. tuberculosis* H37Rv PPIN. We find that *M. tuberculosis* H37Rv proteins involved in the host–pathogen PPIN also tend to have much higher degrees than proteins in the intra-species *M. tuberculosis* H37Rv PPIN. This shows that pathogen proteins involved in the host–pathogen PPIN are also more likely to be hubs in their own intra-species PPIN.

This makes sense as pathogen proteins that interact with human proteins may also have very important functions in the pathogen's own metabolism, and the interaction between hub pathogen proteins with host proteins may be important to switching the pathogen status from managing its own "free-living" metabolism to an "infection-oriented" metabolism.

5.4 Discussion

5.4.1 Homology-based prediction

The homology-based approach for predicting the conserved intra-species PPIs across closely related species was reported more than a decade ago(Matthews et al., 2001), with the assumption that the interaction between a pair of proteins in one species is expected to be conserved in related species. It has also been widely used in predicting inter-species PPIs(Lee et al., 2008; Krishnadev and Srinivasan, 2008; Tyagi et al., 2009; Krishnadev and Srinivasan, 2011; Wuchty, 2011).

However, the limitation of the conventional homology-based approach for predicting inter-species(host–pathogen) PPIs have not been fully discussed. In particular, when

applying this approach in predicting eukaryote-prokaryote PPIs, (i) the differences between eukaryotic and prokaryotic proteins and (ii) the differences between inter-species and intra-species PPI interfaces may all contribute to the limited performance of the conventional homology-based prediction approach in predicting eukaryote-prokaryote host–pathogen PPIs. Therefore, our proposed accurate homology-based prediction approach has merits in overcoming the above two limitations, and should be suitable for predicting eukaryote-prokaryote host–pathogen PPIs in many host–pathogen systems. The main limitation of our accurate homology-based approach lies in the fact that there is a limited amount of source eukaryote-prokaryote PPIs available currently. However, with the rapid advance in technology and the community's increasing interest on host-microbe interaction studies, the eukaryote-prokaryote template PPIs will be much more abundant in the future. This should greatly facilitate the massive application of our accurate prediction approach to many host–pathogen systems in the future.

As a matter of fact, our accurate homology-based approach may not only have merits in predicting eukaryote-prokaryote PPIs, but also can be extended to many other types of inter-species PPI prediction, including eukaryote-archea PPIs(human microbiome; especially in human gut), eukaryote-virus PPIs(e.g. Human-HIV PPIs, Human-EBV PPIs and so one), etc. This can be especially meaningful for predicting human-virus PPIs because (i) there are large differences between human and virus proteins, (ii) human-virus PPI interfaces are also very different from intra-species PPI interfaces, and (iii) abundant template human-virus PPIs are available.

5.4.2 Cancer pathways and enrichment analysis

In several host–pathogen interaction studies, when analyzing the pathway enrichment of host–pathogen PPIN targeted human proteins, cancer-related pathways also show up in the list of most enriched pathways(Evans et al., 2009). According to our study, the presence of cancer pathways makes sense, as cancer shares many similarities with

Pathways	Infection related pathways	Pathways in cancer
Gene No.	1082	326
Overlap between Pathways in cancer and Infection related pathways	169	
Hum-Mtb targeted Human proteins Overlap with HP-PPI targeted Human proteins	204	29
Overlap of the three datasets	20	

Table 5.13: Gene content of cancer pathways and *M. tuberculosis* infection related pathways. This table summarizes the gene content of cancer pathways and *M. tuberculosis* infection related Pathways. We choose one large representative cancer pathway—"Pathways in cancer". The *M. tuberculosis* infection related pathways("infection-related pathways" for short) are: "Focal adhesion, "Proteasome", "Antigen processing and presentation", "MAPK signaling pathway", "Endocytosis", "T cell receptor signaling pathway", "Spliceosome", "Apoptosis", and "Tuberculosis". Hum-Mtb: predicted *H. sapiens–M. tuberculosis* H37Rv PPIN.

pathogen infection, including evading immune response, inducing apoptosis, metastasis and invading the cells. Therefore many essential pathways that are highly interconnected to *M. tuberculosis* infection are also closely related to cancer pathways. These essential pathways are "Apoptosis", "MAPK signaling pathway", "Jak-STAT signaling pathway", "Focal adhesion", etc. On the other hand, cancer pathways may also be an artifact because a substantial number of proteins are in the overlap between the cancer-related pathways and the essential pathways. We conduct some experiments to test this hypothesis. We group all the essential pathways that are related to *M. tuberculosis* infection, and name the collection "infection-related pathways". The collection includes the following pathways, "Focal adhesion", "Proteasome", "Antigen processing and presentation", "MAPK signaling pathway", "Endocytosis", "T cell receptor signaling pathway", "Spliceosome", "Apoptosis", "Tuberculosis". We also choose one large representative cancer pathway ("Pathways in cancer"). We then test the overlap of these two collections of pathways. The results of the analysis are summarized in Table 5.13.

From the results we can see that among the 1082 proteins in "infection-related pathways" and the 326 proteins in "Pathways in cancer", there are 169 proteins over-

lapping between the two datasets. However, the *H. sapiens–M. tuberculosis* H37Rv PPIN predicted by the accurate homology-based prediction approach involves 755 *H. sapiens* proteins. This 755 *H. sapiens* proteins covers 204 proteins in "infection-related pathways" and covers 29 proteins in the "Pathways in cancer". Among these 204 and 29 proteins, 20 of them overlap with each other. This significantly demonstrates our hypothesis that the cancer-related pathways are enriched due to the substantial overlap (in protein members) with infection-related pathways(p-value $\leq$ 1.82E-6).

5.4.3 Impact and possible application of the illuminated sequence and topological properties

Among the key contributions of this work are the discoveries of sequence and topological properties of the proteins involved in the host–pathogen PPIN. Based on the analysis of our predicted host–pathogen PPINs, we see that both host and pathogen proteins involved in host–pathogen PPINs tend to have longer primary sequences, tend to have more domains, tend to have lower charge and tend to be more hydrophilic than proteins in intra-species PPINs. We also see that not only host proteins but also pathogen proteins involved in host–pathogen PPINs tend to be hubs in their own intra-species PPINs.

These important properties maybe useful in application to host–pathogen interaction studies. For example, for assessing the quality of newly predicted or experimentally derived host–pathogen PPIs, we can specifically analyze the sequence and topological properties (primary protein sequences, number of domains, hydrophobicity, charge, domain degrees and intra-species PPIN degrees) of the host and pathogen proteins involved in the host–pathogen PPIs to see how likely the host–pathogen PPIN is valid. These will open more doors for the analysis and assessment of host–pathogen PPINs.

5.5 Conclusion

In this Chapter we have proposed an accurate homology-based approach for predicting host–pathogen PPIs. Our approach specifically overcomes the limitation of the conventional homology-based approach by taking into account two important factors: (i) differences between eukaryotic and prokaryotic proteins, and (ii) differences between intra-species and inter-species PPI interfaces.

Using this accurate homology-based approach, we have predicted 1005 *H. sapiens–M. tuberculosis* H37Rv PPIs. Pathway enrichment analysis, functional enrichment analysis, disease-related gene list enrichment analysis, etc. all support the validity of our prediction results and show that our accurate homology-based approach has better performance in predicting *H. sapiens–M. tuberculosis* H37Rv PPIs than the conventional homology-based approach. The *H. sapiens–M. tuberculosis* H37Rv PPI dataset predicted by our accurate homology-based approach can be used as an important reference for a variety of related studies on *H. sapiens–M. tuberculosis* H37Rv interactions, *M. tuberculosis* H37Rv infections and infectious disease prevention.

We have further analyzed the sequence and topological properties of both the *H. sapiens* and *M. tuberculosis* H37Rv proteins involved in *H. sapiens–M. tuberculosis* H37Rv PPIs. Analysis of sequence properties shows that, both host and pathogen proteins involved in host–pathogen PPIN tend to have longer primary sequences, tend to have more domains, tend to be more hydrophilic and tend to be less positively charged compared to other proteins in intra-species PPIN. Analysis of topological properties shows that not only host proteins but also pathogen proteins involved in the host–pathogen PPIN tend to be hubs in their own intra-species PPIN.

The prediction approach we discussed in this Chapter has merits in applying to many other host–pathogen systems, and the properties that we have discovered through sequence and topological analyses may be helpful in understanding the host–pathogen PPIN and also provide alternative ways to assess predicted host–pathogen PPIN in a

variety of different situations.

Chapter 6

Closing Remarks

6.1 Recap of work done

Using *H. sapiens–M. tuberculosis* H37Rv as a model, we have conducted a systematic computational study on host–pathogen PPIs. We have identified important reliable pathogen PPI datasets that can be used for the analysis of host–pathogen PPIs studies in Chapter 2, which also enables the topological and sequence properties analysis for the first time on the pathogen side. We have developed one of currently most comprehensive pathway databases—IntPath—in Chapter 3, which for the first time enables functional analysis on both host and pathogen. We have proposed two prediction approaches: the accurate DDI-based prediction approach in Chapter 4 and the accurate homology-based prediction approach in Chapter 5. Both of these prediction approaches have better performance than their conventional counterparts. We have provided high-quality predicted *H. sapiens–M. tuberculosis* H37Rv PPIs datasets that can be used for a variety of purposes. Based on the predicted *H. sapiens–M. tuberculosis* H37Rv PPIs, we have observed some very important properties for both host and pathogen proteins involved in the host–pathogen PPIs.

In Chapter 2, we have observed the strikingly low agreement between *M. tuberculosis*

H37Rv B2H and STRING PPI datasets. We have demonstrated the two main causes of this low level of agreement: the H37Rv B2H dataset being of very low quality and the STRING PPI datasets containing many non-direct interactions. The H37Rv B2H dataset is the first—and currently, only—available large-scale experimental protein interaction data on *M. tuberculosis*. It is disappointing that it is of poor quality, and it should not be used in its entirety as a reference *M. tuberculosis* PPI dataset. Nevertheless, we have identified four subsets of this B2H PPI dataset that are more reliable, and can collectively serve as a suitable reference H37Rv physical interaction dataset for many applications. As for the STRING dataset, it is also very noisy. Fortunately, the STRING score is useful for indicating which STRING PPIs have higher quality. We suggest a STRING score threshold set around 770.

In Chapter 3 we overcome the five limitations of current pathway databases that hamper effective use of pathway information. We solve the problem of incompatible data formats in different databases by extracting the pathway-gene and pathway-gene pair relationships. The limitations of inconsistent molecular representations and inconsistent molecular relationship representations have been overcome by our normalization of the data into common gene name representations and common relationship types which are compatible with other database. The problems of inconsistent referrals to pathway names and incomprehensive data from different databases have been solved by the integration of pathway-gene and pathway-gene pair relationships into a unified and comprehensive data source. We achieve compatible data formats, consistent molecular representations, consistent relationship representations, consistent referrals to pathway names and comprehensive data in our IntPath database for several organisms—viz., *H. sapiens*, *S. cerevisiae*, *M. musculus* and *M. tuberculosis* H37Rv. IntPath can maintain a regular update in these organisms and, the methodology we describe here can be applied to other organisms straightforwardly. We believe IntPath will not only facilitate convenient access of the integrated pathway gene relationship data for model organisms

and important pathogens but also greatly boost data analysis and application to the host–pathogen PPIs studies, as demonstrated in Chapter 4 and Chapter 5.

In Chapter 4, we have proposed an stringent DDI-based prediction approach based on high sequence similarity between template domain instances and query domain instances. The assessment based on gold-standard *H. sapiens* PPIs and informative GO annotation shows that the stringent DDI-based approach performs better than the conventional DDI-based approach—its precision is 3–8 times better at the same or even higher levels of recall. This shows, for the first time, that while the conventional idea of DDI mediating PPI is sound, it is critical to apply it using carefully aligned domain instances and checking whether the interaction interfaces are conserved. We have predicted a small set of accurate *H. sapiens–M. tuberculosis* H37Rv PPIs. Through cellular compartment distribution, functional enrichment, and pathway enrichment analysis, we have demonstrated that this small set of accurate *H. sapiens–M. tuberculosis* H37Rv PPIs is valid and closely corresponds to *M. tuberculosis* H37Rv infection. This dataset of *H. sapiens–M. tuberculosis* H37Rv PPIs can be used for a variety of related studies as an important reference.

In Chapter 5, we have proposed an accurate homology-based approach for predicting host–pathogen PPIs. Our approach specifically overcomes the limitation of the conventional homology-based approach—which has relied solely on intra-species PPIs as template—by relying on eukaryote-prokaryote interspecies PPIs as template. That is, by taking into account two important factors: (i) differences between eukaryotic and prokaryotic proteins, and (ii) differences between intra-species and inter-species PPI interfaces. Using this accurate homology-based approach, we have predicted 1005 *H. sapiens–M. tuberculosis* H37Rv PPIs. Thus, for the first time, we show that 3 times more host-pathogen PPIs can be predicted using inter-species template PPIs than using intra-species template PPIs, while requiring 10 times less template PPIs. Moreover, pathway enrichment analysis, functional enrichment analysis, disease-related gene list

enrichment analysis, etc. all support the validity of our prediction results and show that our accurate homology-based approach has better performance in predicting *H. sapiens–M. tuberculosis* H37Rv PPIs than the conventional homology-based approach. We compared the *H. sapiens–M. tuberculosis* H37Rv PPI datasets predicted by my stringent DDI-based approach (92 PPIs) from Chapter 4 and accurate homolgy-based approach (1105 PPIs) from Chapter 5, none of the PPIs overlap between the two datasets. Because the two PPI datasets are predicted by very different methods and based on very different source data, they cover very different proteins. Therefore it is not surprising that the two PPI datasets do not overlap with each other.

The systematic computational studies on *H. sapiens–M. tuberculosis* H37Rv PPIs may serve as an important reference to host–pathogen interaction studies and the methodologies and technologies described in this thesis can be used for many host–pathogen systems.

6.2 Future work

Computational study of host–pathogen interactions is very important for us to have a better understanding on the interplay between pathogen and host, which is crucial for providing better treatment and prevention of infectious diseases. There are many interesting on-going topics in host–pathogen interaction studies that deserve more attention.

We discussed functional analysis based on pathway enrichment using IntPath which captures important functional relevance between host–pathogen PPIs. More powerful pathway analysis tools that can provide a deeper analysis on the connection between host and pathogen pathways and their functional inter-connections should greatly facilitate the understanding of host–pathogen interaction through functional aspects.

The accurate prediction approaches described in Chapter 4 and Chapter 5 can be applied to many host–pathogen systems. This can provide very important host–

pathogen PPI data for many host–pathogen systems which will be crucial for a variety of different studies. Centralized data repositories hosting the predicted host–pathogen PPI data can be developed for convenient access for the community.

As for host–pathogen PPI prediction approaches, integrative approaches can be applied. By integrating related data (e.g. expression, homology, structure, pathways, topology,) in the prediction approaches, better performance may be achieved. In developing integrative prediction approaches, machine learning techniques may be useful.

Without doubt, the progress in computational studies of host–pathogen interactions is already very exciting and there will be lots of more interesting discoveries and breakthroughs in the future.

Bibliography

Anes, E., Kühnel, M., Bos, E., Moniz-Pereira, J., Habermann, A., Griffiths, G., et al. (2003). Selected lipids activate phagosome actin assembly and maturation resulting in killing of pathogenic mycobacteria. *Nature Cell Biology*, 5(9):793–802.

Balakrishnan, S., Tastan, O., Carbonell, J., and Klein-Seetharaman, J. (2009). Alternative paths in HIV-1 targeted human signal transduction pathways. *BMC Genomics*, 10(Suppl 3):S30.

Bateman, A., Coin, L., Durbin, R., Finn, R., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E., et al. (2004). The pfam protein families database. *Nucleic Acids Research*, 32(suppl 1):D138–D141.

Bauer, S., Grossmann, S., Vingron, M., and Robinson, P. (2008). Ontologizer 2.0—a multifunctional tool for go term enrichment analysis and data exploration. *Bioinformatics*, 24(14):1650–1651.

Beißbarth, T. and Speed, T. (2004). GOstat: Find statistically overrepresented gene ontologies within a group of genes. *Bioinformatics*, 20(9):1464–1465.

Black, D. and Bliska, J. (1997). Identification of p130cas as a substrate of yersinia yoph (yop51), a bacterial protein tyrosine phosphatase that translocates into mammalian cells and targets focal adhesions. *The EMBO Journal*, 16(10):2730–2744.

Boyle, E., Weng, S., Gollub, J., Jin, H., Botstein, D., Cherry, J., and Sherlock, G.

(2004). GO:: TermFinder—open source software for accessing gene ontology information and finding significantly enriched gene ontology terms associated with a list of genes. *Bioinformatics*, 20(18):3710–3715.

Brass, A., Dykxhoorn, D., Benita, Y., Yan, N., Engelman, A., Xavier, R., Lieberman, J., and Elledge, S. (2008). Identification of host proteins required for HIV infection through a functional genomic screen. *Science*, 319(5865):921–926.

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

Butler, D. (2000). New fronts in an old war. *Nature*, 406(6797):670–672.

Calderwood, M., Venkatesan, K., Xing, L., Chase, M., Vazquez, A., Holthaus, A., Ewence, A., Li, N., Hirozane-Kishikawa, T., Hill, D., et al. (2007). Epstein–Barr virus and virus human protein interaction maps. *Proceedings of the National Academy of Sciences USA*, 104(18):7606–7611.

Chatr-aryamontri, A., Ceol, A., Peluso, D., Nardozza, A., Panni, S., Sacco, F., Tinti, M., Smolyar, A., Castagnoli, L., Vidal, M., et al. (2009). VirusMINT: a viral protein interaction database. *Nucleic Acids Research*, 37(suppl 1):D669–D673.

Chaussabel, D., Semnani, R., McDowell, M., Sacks, D., Sher, A., and Nutman, T. (2003). Unique gene expression profiles of human macrophages and dendritic cells to phylogenetically distinct parasites. *Blood*, 102(2):672–681.

Chen, L., Wu, L., Wang, Y., and Zhang, X. (2006). Inferring protein interactions from experimental data by association probabilistic method. *PROTEINS: Structure, Function, and Bioinformatics*, 62(4):833–837.

Chertova, E., Chertov, O., Coren, L., Roser, J., Trubey, C., Bess Jr, J., Sowder II, R., Barsov, E., Hood, B., Fisher, R., et al. (2006). Proteomic and biochemical

analysis of purified Human Immunodeficiency Virus Type 1 produced from infected monocyte-derived macrophages. *Journal of Virology*, 80(18):9039–9052.

Chow, E., Razani, B., and Cheng, G. (2007). Innate immune system regulation of nuclear hormone receptors in metabolic diseases. *Journal of Leukocyte Biology*, 82(2):187–195.

Chua, H. N., Sung, W.-K., and Wong, L. (2006). Exploiting indirect neighbours and topological weight to predict protein function from protein–protein interactions. *Bioinformatics*, 22(13):1623–1630.

Cliff, J. M., Lee, J.-S., Constantinou, N., Cho, J.-E., Clark, T. G., Ronacher, K., King, E. C., Lukey, P. T., Duncan, K., Van Helden, P. D., et al. (2013). Distinct phases of blood gene expression pattern through tuberculosis treatment reflect modulation of the humoral immune response. *Journal of Infectious Diseases*, 207(1):18–29.

Cusick, M., Yu, H., Smolyar, A., Venkatesan, K., Carvunis, A., Simonis, N., Rual, J., Borick, H., Braun, P., Dreze, M., et al. (2008). Literature-curated protein interaction datasets. *Nature Methods*, 6(1):39–46.

Davis, F. and Sali, A. (2005). PIBASE: A comprehensive database of structurally defined protein interfaces. *Bioinformatics*, 21(9):1901–1907.

Davis, F. P., Barkan, D. T., Eswar, N., McKerrow, J. H., and Sali, A. (2007). Host-pathogen protein interactions predicted by comparative modeling. *Protein Science*, 16(12):2585–2596.

Demir, E., Cary, M. P., Paley, S., Fukuda, K., et al. (2010). The BioPAX community standard for pathway data sharing. *Nature Biotechnology*, 28(9):935–942.

Dennis Jr, G., Sherman, B., Hosack, D., Yang, J., Gao, W., Lane, H., Lempicki, R., et al. (2003). DAVID: Database for annotation, visualization, and integrated discovery. *Genome Biology*, 4(5):P3.

Doolittle, J. M. and Gomez, S. M. (2010). Structural similarity-based predictions of protein interactions between HIV-1 and *Homo sapiens*. *Virology Journal*, 7(1):82.

Doolittle, J. M. and Gomez, S. M. (2011). Mapping protein interactions between Dengue virus and its human and insect hosts. *PLoS Neglected Tropical Diseases*, 5(2):e954.

Driscoll, T., Dyer, M., Murali, T., and Sobral, B. (2009). PIG–the pathogen interaction gateway. *Nucleic Acids Research*, 37(suppl 1):D647–D650.

Driscoll, T., Gabbard, J., Mao, C., Dalay, O., Shukla, M., Freifeld, C., Hoen, A., Brownstein, J., and Sobral, B. (2011). Integration and visualization of host–pathogen data related to infectious diseases. *Bioinformatics*, 27(16):2279–2287.

Dunphy, K. Y., Senaratne, R. H., Masuzawa, M., Kendall, L. V., and Riley, L. W. (2010). Attenuation of *mycobacterium tuberculosis* functionally disrupted in a fatty acyl-coenzyme a synthetase gene fadd5. *Journal of Infectious Diseases*, 201(8):1232–1239.

Dyer, M., Murali, T., and Sobral, B. (2007). Computational prediction of host-pathogen protein–protein interactions. *Bioinformatics*, 23(13):i159–i166.

Dyer, M., Murali, T., and Sobral, B. (2008). The landscape of human proteins interacting with viruses and other pathogens. *PLoS Pathogens*, 4(2):e32.

Dyer, M., Neff, C., Dufford, M., Rivera, C., Shattuck, D., Bassaganya-Riera, J., Murali, T., and Sobral, B. (2010). The human-bacterial pathogen protein interaction networks of *Bacillus anthracis*, *Francisella tularensis*, and *Yersinia pestis*. *PLoS ONE*, 5(8):e12089.

Dyer, M. D., Murali, T. M., and Sobral, B. W. (2011). Supervised learning and prediction of physical interactions between human and HIV proteins. *Infection, Genetics and Evolution*, 11(5):917–923.

Eddy, S. (2011). Accelerated profile HMM searches. *PLoS Computational Biology*, 7(10):e1002195.

Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792–1797.

Edwards, R., Davey, N., and Shields, D. (2007). SLiMFinder: A probabilistic method for identifying over-represented, convergently evolved, short linear motifs in proteins. *PLoS ONE*, 2(10):e967.

Elliott, B., Kirac, M., Cakmak, A., Yavas, G., Mayes, S., Cheng, E., Wang, Y., Gupta, C., Ozsoyoglu, G., and Ozsoyoglu, M. (2008). PathCase: Pathways database system. *Bioinformatics*, 24(21):2526–2533.

Esposito, C., Marasco, D., Delogu, G., Pedone, E., and Berisio, R. (2011). Heparin-binding hemagglutinin hbha from *mycobacterium tuberculosis* affects actin polymerisation. *Biochemical and Biophysical Research Communications*, 410(2):339–344.

Evans, P., Dampier, W., Ungar, L., and Tozeren, A. (2009). Prediction of HIV-1 virus-host protein interactions using virus and host sequence motifs. *BMC Medical Genomics*, 2(1):27.

Evsikov, A., Dolan, M., Genrich, M., Patek, E., and Bult, C. (2009). MouseCyc: A curated biochemical pathways database for the laboratory mouse. *Genome Biology*, 10(8):R84.

Fahey, M., Bennett, M., Mahon, C., Jäger, S., Pache, L., Kumar, D., Shapiro, A., Rao, K., Chanda, S., Craik, C., et al. (2011). GPS-Prot: A web-based visualization platform for integrating host-pathogen interaction data. *BMC Bioinformatics*, 12(1):298.

Finn, R., 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.

Franzosa, E. and Xia, Y. (2011). Structural principles within the human-virus protein-protein interaction network. *Proceedings of the National Academy of Sciences USA*, 108(26):10538–10543.

Fu, W., Sanders-Beer, B., Katz, K., Maglott, D., Pruitt, K., and Ptak, R. (2009). Human immunodeficiency virus type 1, human protein interaction database at NCBI. *Nucleic Acids Research*, 37(suppl 1):D417–D422.

Gagliardi, M., Teloni, R., Giannoni, F., Mariotti, S., Remoli, M., Sargentini, V., Videtta, M., Pardini, M., De Libero, G., Coccia, E., et al. (2009). Mycobacteria exploit p38 signaling to affect cd1 expression and lipid antigen presentation by human dendritic cells. *Infection and Immunity*, 77(11):4947–4952.

Gilbert, D. (2005). Biomolecular interaction network database. *Briefings in Bioinformatics*, 6(2):194–198.

Gillespie, J., Wattam, A., Cammer, S., Gabbard, J., Shukla, M., Dalay, O., Driscoll, T., Hix, D., Mane, S., Mao, C., et al. (2011). PATRIC: The comprehensive bacterial bioinformatics resource with a focus on human pathogenic species. *Infection and Immunity*, 79(11):4286–4298.

Global Tuberculosis Programme, W. H. O. (2010). *Global Tuberculosis Control: WHO Report*. Global Tuberculosis Programme, World Health Organization.

Goh, W., Fan, M., Low, H. S., Sergot, M., Wong, L., et al. (2013). Enhancing the utility of proteomics signature profiling (psp) with pathway derived subnets (pdss), performance analysis and specialised ontologies. *BMC genomics*, 14(1):35.

Goh, W., Lee, Y. H., Zubaidah, R. M., Jin, J., Dong, D., Chung, M. C. M., and Wong, L. (2011). Network-based pipeline for analyzing MS data: An application toward liver cancer. *Journal of Proteome Research*, 10(5):2261–2272.

Grigoriev, A. (2001). A relationship between gene expression and protein interactions on the proteome scale: Analysis of the bacteriophage T7 and the yeast *saccharomyces cerevisiae*. *Nucleic Acids Research*, 29(17):3513–3519.

Guérin, I. and de Chastellier, C. (2000a). Disruption of the actin filament network affects delivery of endocytic contents marker to phagosomes with early endosome characteristics: the case of phagosomes with pathogenic mycobacteria. *European Journal of Cell Biology*, 79(10):735–749.

Guérin, I. and de Chastellier, C. (2000b). Pathogenic mycobacteria disrupt the macrophage actin filament network. *Infection and Immunity*, 68(5):2655–2662.

Güldener, U., Münsterkötter, M., Oesterheld, M., Pagel, P., Ruepp, A., Mewes, H., and Stümpflen, V. (2006). MPact: The MIPS protein interaction resource on yeast. *Nucleic Acids Research*, 34(suppl 1):D436–D441.

Haider, S., Ballester, B., Smedley, D., Zhang, J., Rice, P., and Kasprzyk, A. (2009). Biomart central portal—unified access to biological data. *Nucleic Acids Research*, 37(Database issue):W23–W27.

Han, D., Kim, H., Jang, W., Lee, S., and Suh, J. (2004). PreSPI: A domain combination based prediction system for protein–protein interaction. *Nucleic Acids Research*, 32(21):6312–6320.

Hawkins, T. and Kihara, D. (2007). Function prediction of uncharacterized proteins. *Journal of bioinformatics and computational biology*, 5(01):1–30.

Hayashida, M., Ueda, N., Akutsu, T., et al. (2004). A simple method for inferring strengths of protein-protein interactions. *Genome Informatics Series*, 15(1):56–68.

Henikoff, S. and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences USA*, 89(22):10915–10919.

Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S., Vingron, M., Roechert, B., Roepstorff, P., Valencia, A., et al. (2004). IntAct: An open source molecular interaction database. *Nucleic Acids Research*, 32(suppl 1):D452–D455.

Hestvik, A., Hmama, Z., and Av-Gay, Y. (2006). Mycobacterial manipulation of the host cell. *FEMS Microbiology Reviews*, 29(5):1041–1050.

Holm, L., Kääriäinen, S., Rosenström, P., and Schenkel, A. (2008). Searching protein structure databases with DaliLite v. 3. *Bioinformatics*, 24(23):2780–2781.

Hugo, W., Ng, S., and Sung, W. (2011). D-SLIMMER: Domain-SLiM Interaction Motifs Miner for sequence based protein-protein interaction data. *Journal of Proteome Research*, 10(12):5285–5295.

Hulo, N., Bairoch, A., Bulliard, V., Cerutti, L., Cuche, B., De Castro, E., Lachaize, C., Langendijk-Genevaux, P., and Sigrist, C. (2008). The 20 years of PROSITE. *Nucleic Acids Research*, 36(suppl 1):D245–D249.

Hunter, S., Jones, P., Mitchell, A., Apweiler, R., Attwood, T. K., Bateman, A., Bernard, T., Binns, D., Bork, P., Burge, S., et al. (2012). Interpro in 2011: New developments in the family and domain prediction database. *Nucleic Acids Research*, 40(D1):D306–D312.

Isserlin, R., El-Badrawi, R. A., and Bader, G. D. (2011). The Biomolecular Interaction Network Database in PSI-MI 2.5. *Database: The Journal of Biological Databases and Curation*, 2011.

Itzhaki, Z., Akiva, E., and Margalit, H. (2010). Preferential use of protein domain pairs as interaction mediators: order and transitivity. *Bioinformatics*, 26(20):2564–2570.

Jamwal, S., Midha, M. K., Verma, H. N., Basu, A., Rao, K. V., and Manivel, V. (2013). Characterizing virulence-specific perturbations in the mitochondrial function of macrophages infected with *mycobacterium tuberculosis*. *Scientific Reports*, 3:1328.

Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P., Schmidt, E., de Bono, B., Jassal, B., Gopinath, G., Wu, G., Matthews, L., et al. (2005). Reactome: A knowledgebase of biological pathways. *Nucleic Acids Research*, 33(suppl 1):D428–D432.

Kamburov, A., Pentchev, K., Galicka, H., Wierling, C., Lehrach, H., and Herwig, R. (2011). ConsensusPathDB: Toward a more complete picture of cell biology. *Nucleic Acids Research*, 39(suppl 1):D712.

Karp, P. D. (2001). Pathway databases: A case study in computational symbolic theories. *Science*, 293(5537):2040–2044.

Karp, P. D., Ouzounis, C. A., Moore-Kochlacs, C., Goldovsky, L., Kaipa, P., Ahrén, D., Tsoka, S., Darzentas, N., Kunin, V., and López-Bigas, N. (2005). Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Research*, 33(19):6083–6089.

Kelder, T., Pico, A. R., Hanspers, K., van Iersel, M. P., Evelo, C., and Conklin, B. R. (2009). Mining biological pathways using WikiPathways web services. *PLoS ONE*, 4(7):e6447.

Kerrien, S., Aranda, B., Breuza, L., Bridge, A., Broackes-Carter, F., Chen, C., Duesbury, M., Dumousseau, M., Feuermann, M., Hinz, U., et al. (2012). The IntAct molecular interaction database in 2012. *Nucleic Acids Research*, 40(D1):D841–D846.

Kim, W., Kim, K., Lee, E., Marcotte, E., Kim, H., and Suh, J. (2007). Identification of disease specific protein interactions between the gastric cancer causing pathogen, *H. pylori*, and human hosts using protein network modeling and gene chip analysis. *BioChip Journal*, 1:179–187.

Kim, W., Park, J., Suh, J., et al. (2002). Large scale statistical prediction of protein-protein interaction by potentially interacting domain (PID) pair. *Genome Informatics Series*, 13:42–50.

König, R., Zhou, Y., Elleder, D., Diamond, T., Bonamy, G., Irelan, J., Chiang, C., Tu, B., De Jesus, P., Lilley, C., et al. (2008). Global analysis of host-pathogen interactions that regulate early-stage HIV-1 replication. *Cell*, 135(1):49–60.

Koul, A., Choidas, A., Treder, M., Tyagi, A., Drlica, K., Singh, Y., and Ullrich, A. (2000). Cloning and characterization of secretory tyrosine phosphatases of *mycobacterium tuberculosis*. *Journal of Bacteriology*, 182(19):5425–5432.

Koul, A., Herget, T., Klebl, B., and Ullrich, A. (2004). Interplay between mycobacteria and host signalling pathways. *Nature Reviews Microbiology*, 2(3):189–202.

Krachler, A., Woolery, A., and Orth, K. (2011). Manipulation of kinase signaling by bacterial pathogens. *The Journal of Cell Biology*, 195(7):1083–1092.

Krishnadev, O. and Srinivasan, N. (2008). A data integration approach to predict host-pathogen protein-protein interactions: Application to recognize protein interactions between human and a malarial parasite. *In Silico Biology*, 8(3):235–250.

Krishnadev, O. and Srinivasan, N. (2011). Prediction of protein-protein interactions between human host and a pathogen and its application to three pathogenic bacteria. *International Journal of Biological Macromolecules*, 48:613–619.

Krishnan, M., Ng, A., Sukumaran, B., Gilfoy, F., Uchil, P., Sultana, H., Brass, A., Adametz, R., Tsui, M., Qian, F., et al. (2008). RNA interference screen for human genes associated with West Nile virus infection. *Nature*, 455(7210):242–245.

Lee, S. A., Chan, C., Tsai, C. H., Lai, J. M., Wang, F. S., Kao, C. Y., and Huang, C. Y. (2008). Ortholog-based protein-protein interaction prediction and its application to inter-species interactions. *BMC Bioinformatics*, 9(Suppl 12):S11.

Lee, W., VanderVen, B. C., Fahey, R. J., and Russell, D. G. (2013). Intracellular *mycobacterium tuberculosis* exploits host-derived fatty acids to limit metabolic stress. *Journal of Biological Chemistry*, 288(10):6788–6800.

Li, Y. and Agarwal, P. (2009). A pathway-based view of human diseases and disease relationships. *PloS one*, 4(2):e4346.

Liu, P., Stenger, S., Li, H., Wenzel, L., Tan, B., Krutzik, S., Ochoa, M., Schauber, J., Wu, K., Meinken, C., et al. (2006). Toll-like receptor triggering of a vitamin d-mediated human antimicrobial response. *Science Signalling*, 311(5768):1770.

Maglott, D., Ostell, J., Pruitt, K. D., and Tatusova, T. (2005). Entrez gene: gene-centered information at ncbi. *Nucleic Acids Research*, 33(Database issue):D54–D58.

Marrero, J., Rhee, K. Y., Schnappinger, D., Pethe, K., and Ehrt, S. (2010). Gluconeogenic carbon flow of tricarboxylic acid cycle intermediates is critical for *mycobacterium tuberculosis* to establish and maintain infection. *Proceedings of the National Academy of Sciences USA*, 107(21):9819–9824.

Matthews, L., Vaglio, P., Reboul, J., Ge, H., Davis, B., Garrels, J., Vincent, S., and Vidal, M. (2001). Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs". *Genome Research*, 11(12):2120–2126.

McGarvey, P., Huang, H., Mazumder, R., Zhang, J., Chen, Y., Zhang, C., Cammer, S., Will, R., Odle, M., Sobral, B., et al. (2009). Systems integration of biodefense omics data for analysis of pathogen-host interactions and identification of potential targets. *PLoS ONE*, 4(9):e7162.

Mishra, G., Suresh, M., Kumaran, K., Kannabiran, N., Suresh, S., Bala, P., Shivakumar, K., Anuradha, N., Reddy, R., Raghavan, T., et al. (2006). Human protein reference database—2006 update. *Nucleic Acids Research*, 34(suppl 1):D411–D414.

Murzin, A., Brenner, S., Hubbard, T., Chothia, C., et al. (1995). SCOP: A structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247(4):536–540.

Navratil, V., De Chassey, B., Meyniel, L., Delmotte, S., Gautier, C., André, P., Lotteau, V., and Rabourdin-Combe, C. (2009). VirHostNet: A knowledge base for the management and the analysis of proteome-wide virus–host interaction networks. *Nucleic Acids Research*, 37(suppl 1):D661–D668.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. (1999). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 27(1):29–34.

Ott, D. (2008). Cellular proteins detected in HIV-1. *Reviews in Medical Virology*, 18(3):159–175.

Pagel, P., Kovac, S., Oesterheld, M., Brauner, B., Dunger-Kaltenbach, I., Frishman, G., Montrone, C., Mark, P., Stümpflen, V., Mewes, H., et al. (2005). The MIPS mammalian protein–protein interaction database. *Bioinformatics*, 21(6):832–834.

Parrish, J. R., Yu, J., Liu, G., Hines, J. A., Chan, J. E., Mangiola, B. A., Zhang, H., Pacifico, S., Fotouhi, F., DiRita, V. J., et al. (2007). A proteome-wide protein interaction map for *campylobacter jejuni*. *Genome biology*, 8(7):R130.

Persson, C., Carballeira, N., Wolf-Watz, H., and Fällman, M. (1997). The ptpase yoph inhibits uptake of yersinia, tyrosine phosphorylation of p130cas and fak, and the associated accumulation of these proteins in peripheral focal adhesions. *The EMBO Journal*, 16(9):2307–2318.

Pico, A. R., Kelder, T., van Iersel, M. P., Hanspers, K., Conklin, B. R., and Evelo, C. (2008). WikiPathways: Pathway editing for the people. *PLoS Biology*, 6(7):e184.

Prieto, C. and De Las Rivas, J. (2006). APID: Agile protein interaction dataanalyzer. *Nucleic Acids Research*, 34(suppl 2):W298–W302.

Pu, S., Wong, J., Turner, B., Cho, E., and Wodak, S. J. (2009). Up-to-date catalogues of yeast protein complexes. *Nucleic Acids Research*, 37(3):825–831.

Puntervoll, P., Linding, R., Gemünd, C., Chabanis-Davidson, S., Mattingsdal, M., Cameron, S., Martin, D., Ausiello, G., Brannetti, B., Costantini, A., et al. (2003). ELM server: A new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic Acids Research*, 31(13):3625–3630.

Qi, Y., Tastan, O., Carbonell, J. G., Klein-Seetharaman, J., and Weston, J. (2010). Semi-supervised multi-task learning for predicting interactions between HIV-1 and human proteins. *Bioinformatics*, 26(18):i645–i652.

Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R., and Lopez, R. (2005). Interproscan: Protein domains identifier. *Nucleic Acids Research*, 33(suppl 2):W116–W120.

Rachman, H., Strong, M., Ulrichs, T., Grode, L., Schuchhardt, J., Mollenkopf, H., Kosmiadi, G., Eisenberg, D., and Kaufmann, S. (2006). Unique transcriptome signature of *Mycobacterium tuberculosis* in pulmonary tuberculosis. *Infection and Immunity*, 74(2):1233–1242.

Ranjit, K. and Bindu, N. (2010). HPIDB—a unified resource for host-pathogen interactions. *BMC Bioinformatics*, 11(Suppl 6):S16.

Rappoport, N. and Linial, M. (2012). Viral proteins acquired from a host converge to simplified domain architectures. *PLoS Computational Biology*, 8(2):e1002364.

Razick, S., Magklaras, G., and Donaldson, I. (2008). iRefIndex: A consolidated protein interaction database with provenance. *BMC Bioinformatics*, 9(1):405.

Romero, P., Wagg, J., Green, M. L., Kaiser, D., Krummenacker, M., and Karp, P. D. (2005). Computational prediction of human metabolic pathways from the complete human genome. *Genome Biology*, 6(1):R2.

Salomonis, N., Hanspers, K., Zambon, A., Vranizan, K., Lawlor, S., Dahlquist, K., Doniger, S., Stuart, J., Conklin, B., and Pico, A. (2007). GenMAPP 2: New features and resources for pathway analysis. *BMC Bioinformatics*, 8(1):217.

Salwinski, L., Miller, C., Smith, A., Pettit, F., Bowie, J., and Eisenberg, D. (2004). The database of interacting proteins: 2004 update. *Nucleic Acids Research*, 32(suppl 1):D449–D451.

Sassetti, C. and Rubin, E. (2003). Genetic requirements for mycobacterial survival during infection. *Proceedings of the National Academy of Sciences USA*, 100(22):12989–12994.

Sato, S., Shimoda, Y., Muraki, A., Kohara, M., Nakamura, Y., and Tabata, S. (2007). A large-scale protein–protein interaction analysis in *Synechocystis* sp. PCC6803. *DNA Research*, 14(5):207–216.

Schaefer, C., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T., and Buetow, K. (2009). PID: The pathway interaction database. *Nucleic Acids Research*, 37(suppl 1):D674–D679.

Seal, R., Gordon, S., Lush, M., Wright, M., and Bruford, E. (2011). genenames.org: The HGNC resources in 2011. *Nucleic Acids Research*, 39(Database issue):D519.

Sergey, K., Mayya, S., Yulia, D., Amarnath, G., Animesh, R., Julia, P., and Michael, B. (2011). BiologicalNetworks—tools enabling the integration of multi-scale data for the host-pathogen studies. *BMC Systems Biology*, 5(1):7.

Sessions, O., Barrows, N., Souza-Neto, J., Robinson, T., Hershey, C., Rodgers, M.,

Ramirez, J., Dimopoulos, G., Yang, P., Pearson, J., et al. (2009). Discovery of insect and human dengue virus host factors. *Nature*, 458(7241):1047–1050.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11):2498–2504.

Shi, L., Sohaskey, C. D., Pfeiffer, C., Datta, P., Parks, M., McFadden, J., North, R. J., and Gennaro, M. L. (2010). Carbon flux rerouting during *mycobacterium tuberculosis* growth arrest. *Molecular Microbiology*, 78(5):1199–1215.

Singh, I., Tastan, O., and Klein-Seetharaman, J. (2010). Comparison of virus interactions with human signal transduction pathways. In *Proceedings of the First ACM International Conference on Bioinformatics and Computational Biology*, pages 17–24. ACM.

Singh, R., Rao, V., Shakila, H., Gupta, R., Khera, A., Dhar, N., Singh, A., Koul, A., Singh, Y., Naseema, M., et al. (2003). Disruption of mptpB impairs the ability of *Mycobacterium tuberculosis* to survive in guinea pigs. *Molecular Microbiology*, 50(3):751–762.

Smoot, M., Ono, K., Ruscheinski, J., Wang, P., and Ideker, T. (2011). Cytoscape 2.8: New features for data integration and network visualization. *Bioinformatics*, 27(3):431–432.

Soh, D., Dong, D., Guo, Y., and Wong, L. (2010). Consistency, comprehensiveness, and compatibility of pathway databases. *BMC Bioinformatics*, 11:449.

Soh, D., Dong, D., Guo, Y., and Wong, L. (2011). Finding consistent disease subnetworks across microarray datasets. *BMC bioinformatics*, 12(Suppl 13):S15.

Sprinzak, E. and Margalit, H. (2001). Correlated sequence-signatures as markers of protein-protein interaction. *Journal of Molecular Biology*, 311(4):681–692.

Stark, C., Breitkreutz, B., Chatr-Aryamontri, A., Boucher, L., Oughtred, R., Livstone, M., Nixon, J., Van Auken, K., Wang, X., Shi, X., et al. (2011). The BioGRID interaction database: 2011 update. *Nucleic Acids Research*, 39(suppl 1):D698–D704.

Stein, A., Céol, A., and Aloy, P. (2011). 3did: Identification and classification of domain-based interactions of known three-dimensional structure. *Nucleic Acids Research*, 39(suppl 1):D718–D723.

Stobbe, M., Houten, S., Jansen, G., van Kampen, A., and Moerland, P. (2011). Critical assessment of human metabolic pathway databases: A stepping stone for future integration. *BMC Systems Biology*, 5(1):165.

Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., Doerks, T., Stark, M., Muller, J., Bork, P., et al. (2011). The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Research*, 39(suppl 1):D561–D568.

Tastan, O., Qi, Y., Carbonell, J. G., and Klein-Seetharaman, J. (2009). Prediction of interactions between HIV-1 and human proteins by information integration. *Pacific Symposium on Biocomputing*, 14:516–527.

The UniProt Consortium (2012). Reorganizing the protein space at the universal protein resource (uniprot). *Nucleic Acids Research*, 40(D1):D71–D75.

Thieu, T., Joshi, S., Warren, S., and Korkin, D. (2012). Literature mining of host–pathogen interactions: comparing feature-based supervised learning and language-based approaches. *Bioinformatics*, 28(6):867–875.

Ting, L., Kim, A., Cattamanchi, A., and Ernst, J. (1999). *Mycobacterium tuberculosis*

inhibits ifn-γ transcriptional responses without inhibiting activation of stat1. *The Journal of Immunology*, 163(7):3898–3906.

Toossi, Z., Xia, L., Wu, M., and Salvekar, A. (1999). Transcriptional activation of hiv by *mycobacterium tuberculosis* in human monocytes. *Clinical and Experimental Immunology*, 117(2):324–330.

Turner, B., Razick, S., Turinsky, A., Vlasblom, J., Crowdy, E., Cho, E., Morrison, K., Donaldson, I., and Wodak, S. (2010). iRefWeb: Interactive analysis of consolidated protein interaction data and their supporting evidence. *Database: The Journal of Biological Databases and Curation*, 2010:baq023.

Tyagi, N., Krishnadev, O., and Srinivasan, N. (2009). Prediction of protein–protein interactions between *Helicobacter pylori* and a human host. *Molecular BioSystems*, 5(12):1630–1635.

Valone, S., Rich, E., Wallis, R., and Ellner, J. (1988). Expression of tumor necrosis factor in vitro by human mononuclear phagocytes stimulated with whole mycobacterium bovis bcg and mycobacterial antigens. *Infection and Immunity*, 56(12):3313–3315.

van Iersel, M., Kelder, T., Pico, A., Hanspers, K., Coort, S., Conklin, B., and Evelo, C. (2008). Presenting and exploring biological pathways with PathVisio. *BMC Bioinformatics*, 9(1):399.

Wallis, R., Fujiwara, H., and Ellner, J. (1986). Direct stimulation of monocyte release of interleukin 1 by mycobacterial protein antigens. *The Journal of Immunology*, 136(1):193–196.

Wang, Y., Cui, T., Zhang, C., Yang, M., Huang, Y., Li, W., Zhang, L., Gao, C., He, Y., Li, Y., et al. (2010). Global Protein–Protein Interaction Network in the Human Pathogen *Mycobacterium tuberculosis* H37Rv. *Journal of Proteome Research*, 9(12):6665–6677.

Winnenburg, R., Baldwin, T., Urban, M., Rawlings, C., Köhler, J., and Hammond-Kosack, K. (2006). PHI-base: A new database for pathogen host interactions. *Nucleic Acids Research*, 34(suppl 1):D459–D464.

Winnenburg, R., Urban, M., Beacham, A., Baldwin, T., Holland, S., Lindeberg, M., Hansen, H., Rawlings, C., Hammond-Kosack, K., and Köhler, J. (2008). PHI-base update: Additions to the pathogen–host interaction database. *Nucleic Acids Research*, 36(suppl 1):D572–D576.

Wong, L. (2011). Using biological networks in protein function prediction and gene expression analysis. *Internet Mathematics*, 7(4):274–298.

Wong, L. and Liu, G. (2010). Protein interactome analysis for countering pathogen drug resistance. *Journal of Computer Science and Technology*, 25(1):124–130.

Wuchty, S. (2011). Computational prediction of host-parasite protein interactions between *P. falciparum* and *H. sapiens*. *PLoS ONE*, 6(11):e26960.

Xiang, Z., Tian, Y., He, Y., et al. (2007). PHIDIAS: A pathogen-host interaction data integration and analysis system. *Genome Biology*, 8(7):R150.

Yadav, M. K., Pandy, S. K., and Swati, D. (2013). Drug target prioritization in plasmodium falciparum through metabolic network analysis, and inhibitor designing using virtual screening and docking approach. *Journal of Bioinformatics and Computational Biology*, 11(2):1350003.

Yellaboina, S., Tasneem, A., Zaykin, D., Raghavachari, B., and Jothi, R. (2011). Domine: A comprehensive collection of known and predicted domain-domain interactions. *Nucleic Acids Research*, 39(suppl 1):D730–D735.

Yeung, M., Houzet, L., Yedavalli, V., and Jeang, K. (2009). A genome-wide short hairpin RNA screening of jurkat T-cells for human proteins contributing to productive HIV-1 replication. *Journal of Biological Chemistry*, 284(29):19463–19473.

Yu, H., Braun, P., Yıldırım, M. A., Lemmens, I., Venkatesan, K., Sahalie, J., Hirozane-Kishikawa, T., Gebreab, F., Li, N., Simonis, N., et al. (2008). High-quality binary protein interaction map of the yeast interactome network. *Science*, 322(5898):104–110.

Zanzoni, A., Montecchi-Palazzi, L., Quondam, M., Ausiello, G., Helmer-Citterich, M., and Cesareni, G. (2002). MINT: A Molecular INTeraction database. *FEBS Letters*, 513(1):135–140.

Zhang, C., Crasta, O., Cammer, S., Will, R., Kenyon, R., Sullivan, D., Yu, Q., Sun, W., Jha, R., Liu, D., et al. (2008). An emerging cyberinfrastructure for biodefense pathogen and pathogen–host data. *Nucleic Acids Research*, 36(suppl 1):D884–D891.

Zhang, M. and Leong, H. (2012). BBH-LS: An algorithm for computing positional homologs using sequence and gene context similarity. *BMC Systems Biology*, 6(Suppl 1):S22.

Zhao, Z., Xia, J., Tastan, O., Singh, I., Kshirsagar, M., Carbonell, J., and Klein-Seetharaman, J. (2011). Virus interactions with human signal transduction pathways. *International Journal of Computational Biology and Drug Design*, 4(1):83–105.

Zhou, C., Smith, J., Lam, M., Zemla, A., Dyer, M., and Slezak, T. (2007). MvirDB a microbial database of protein toxins, virulence factors and antibiotic resistance genes for bio-defence applications. *Nucleic Acids Research*, 35(suppl 1):D391–D394.

Zhou, H., Jin, J., and Wong, L. (2013). Progress in computational studies of host-pathogen interactions. *Journal of Bioinformatics and Computational Biology*, 11(2):1230001.

Zhou, H., Jin, J., Zhang, H., Bo, Y., Wozniak, M., and Wong, L. (2012). Intpath—an integrated pathway gene relationship database for model organisms and important pathogens. *BMC Systems Biology*, 6(Suppl 2):S2.

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

Zhou, H., Xu, M., Huang, Q., Gates, A., Zhang, X., Castle, J., Stec, E., Ferrer, M., Strulovici, B., Hazuda, D., et al. (2008). Genome-scale RNAi screen for host factors required for HIV replication. *Cell Host & Microbe*, 4(5):495–504.

Appendix A

Additional Files

This appendix contains the additional files for the Chapter 2, 4 and 5.

A.1 Additional file 1 — Reliable *M. tuberculosis* H37Rv B2H PPI datasets

We identified the reliable M. tuberculosis H37Rv B2H PPI datasets in Chapter 2, list in four text files, tab delimited.

A.2 Additional file 2 — Predicted *H.sapiens-M. tuberculosis* H37Rv PPI datasets

We predicted *H.sapiens-M. tuberculosis* H37Rv PPIs using our accurate DDI-based prediction approach in Chapter 4. The predicted PPI data are recorded in simple text format in additional file 2.

A.3 Additional file 3 — Predicted *H. sapiens-M. tuberculosis* H37Rv PPI datasets

We predicted 1005 *H. sapiens-M. tuberculosis* H37Rv PPIs using the accurate homology-based prediction approach in Chapter 5. All the PPI data are recorded in simple text format in this additional file.