Computational prediction of host-pathogen protein-protein interactions

Ibrahim H.I. Ahmed

Supervisor-: Professor Alan Christoffels

Cosupervisor-: Professor Peter Witbooi

Dissertation presented in fulfillment of the requirements for the degree of

Doctor of Philosophy

at the South African National Bioinformatics Institute

Faculty of Natural Sciences

University of the Western Cape

March 20, 2017

Abstract

Computational prediction of host-pathogen protein-protein interactions

I.H.I Ahmed (Ibrahim)

Supervised machine learning approaches have been applied successfully to the prediction of protein-protein interactions (PPIs) within a single organism, i.e., intra-species predictions. However, because of the absence of large amounts of experimentally validated PPIs data for training and testing, fewer studies have successfully applied these techniques to host-pathogen PPI, i.e., inter-species comparisons. Among the host-pathogen studies, most of them have focused on human-virus interactions and specifically human-HIV PPI data. Additional improvements to machine learning techniques and feature sets are important to improve the classification accuracy for host-pathogen protein-protein interactions prediction.

The primary aim of this bioinformatics thesis was to develop a binary classifier with an appropriate feature set for host-pathogen protein-protein interaction prediction using published human-*Hepatitis C virus* PPI, and to test the model on available host-pathogen data for human-*Bacillus anthracis* PPI. Twelve different feature sets were compared to find the optimal set.

The feature selection process reveals that our novel quadruple feature (a subsequence of four consecutive amino acid) combined with sequence similarity and human interactome network properties (such as degree, cluster coefficient, and betweenness centrality) were

the best set. The optimal feature set outperformed those in the relevant published material, giving 95.9% sensitivity, 91.6% specificity and 89.0% accuracy.

Using our optimal features set, we developed a neural network model to predict PPI between human-*Mycobacterium tuberculosis*. The strategy is to develop a model trained with intra-species PPI data and extend it to inter-species prediction. However, the lack of experimentally validated PPI data between human-Mycobacterium tuberculosis (M-tuberculosis), leads us to first assess the feasibility of using validated intra-species PPI data to build a model for inter-species PPI. In this model we used human intra-species PPI combined with *Bacillus anthracis* intra-species data to develop a binary classification model and extend the model for human-*Bacillus anthracis* inter-species prediction. Thus, we test our hypotheses on known human-*Bacillus anthracis* PPI data and the result shows good performance with 89.0% as average accuracy.

The same approach was extended to the prediction of PPI between human-*Mycobacterium tuberculosis*. The predicted human-*M-tuberculosis* PPI data were further validated using functional enrichment of experimentally verified secretory proteins in *M-tuberculosis*, cellular compartment analysis and pathway enrichment analysis. Results show that five of the *M-tuberculosis* secretory proteins within an infected host macrophage that correspond to the mycobacterial virulent strain H37Rv were extracted from the human-*M-tuberculosis* PPI dataset predicted by our model. Finally, a web server was created to predict PPIs between human and *Mycobacterium tuberculosis* which is available online at URL:http://hppredict.sanbi.ac.za.

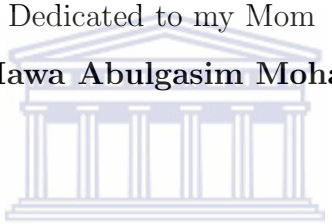
In summary, the concepts, techniques and technologies developed as part of this thesis have the potential to contribute not only to the understanding PPI analysis between human and *Mycobacterium tuberculosis*, but can be extended to other pathogens. Further materials related to this study are available at ftp://ftp.sanbi.ac.za/machine learning.

Keywords: Machine learning, feature selection, web server, support vector machine, artificial neural network, *Mycobacterium tuberculosis*, *Bacillus anthracis*.

Dedication

Dedicated to my Mom

Hawa Abulgasim Mohammed



Declaration

I declare that *Computational prediction of host-pathogen protein-protein interactions* is my own work, that it has not been submitted for any degree or examination in any other university, and that all the sources I have used or quoted have been indicated and acknowledged by complete references.

Ibrahim I.H.I. Ahmed July, 2016

Signed: .....

Acknowledgements

I would like to express my deep and sincere gratitude to my supervisor Professor Alan Christoffels and co supervisor Professor Peter Joseph Witbooi for the supervision of this dissertation, their guidance, encouragement and patience. I wish to thank my siblings and my entire extended family for providing a loving environment for me. Most importantly, I wish to thank my beloved parents, Hussin Ibrahim and Hawa Abolgasim for their care and love. Lastly, my special gratitude goes to SANBI and the National Research Fund (NRF) for

financial support during this doctoral research.

Publications Pending

Ibrahim Ahmed, Peter Witbooi, Eileen Hoal, Alan Christoffels. **Features selection and its application on host-pathogen PPI prediction**. In preparation.

Ibrahim Ahmed, Peter Witbooi, Eileen Hoal, Alan Christoffels. **Human-***Mycobacterium tuberculosis* **PPI Prediction**. In preparation.

Ibrahim Ahmed, Peter Witbooi, Eileen Hoal, Alan Christoffels. **HPPredict: Online Human-***M-tuberculosis* **PPI Predictor.** In preparation.

List of Acronyms

PPI Protein-Protein Interaction

APID Agile Protein Interaction Data Analyzer

DNA Deoxyribonucleic acid

PHIDIAS PathogenHost Interaction Data Integration and Analysis System

RCBPR Resource Center for Biodefense Proteomics Research

PIG Pathogen Interaction Gateway

PATRIC The Pathosystems Resource Integration Center

BIND Biomolecular Interaction Network Database

BioGRID Biological General Repository for Interaction Datasets

MIPS The Munich Information Center for Protein Sequences

MINT The Molecular INTeraction database

ROC The receiver operating characteristic curve

GO Gene ontology

Genome The complete set of genes of an organism

Proteome The complete set of proteins expressed by a genome

Y2H Yeast two-hybrid

TAP-MS Tandem affinity purification mass spectrometry

G(V,E) Graph where V is the set of nodes and E is the set of edges

K_v Degree of vertex v

C_v Clustering coefficient of vertex v

B_v Betweenness centrality of vertex v

Contents

List of Figures

Chapter 1

Introduction

The ability of cells to sense their surrounding environment and respond in an appropriate manner is essential for the normal functioning of every living organism. Cells are constantly exposed to numerous stimuli and respond accordingly. These correct responses are based on numerous intracellular signaling networks that are mostly achieved by proteins. Proteins are the building blocks that facilitate most biological processes in a cell, including cell growth, proliferation, nutrient uptake, gene expression, morphology, inter cellular communication, apoptosis and motility. A protein can be expected to work in relative isolation, but the majority are expected to operate in accordance with other proteins in complexes and networks to regulate a myriad of processes that impact on cellular structure and function, (Herbert and Hethcote, 2000). Some of these processes include cell-cycle control, differentiation, protein folding, signaling, transcription, translation, post-translational modification and transport. Most of these processes can be achieved through protein-protein interactions (PPI).

Protein-protein interactions refer to physical contacts established between two or more proteins as a results of biochemical events or electrostatic forces. Therefore, PPIs and their associated networks are intrinsic to understanding cellular processes, such as enzymatic activity, immunological recognition, DNA repair, network pathway, signaling cascades and transcription control. On the other hand, many human diseases can be traced to aberrant protein-protein interactions (Sandeep et al., 2010), and include endogenous

proteins (Moller and Hoal, 2010), proteins from pathogens or both (Schluger and Rom, 1998). However, unraveling physical interaction between two proteins is essential for understanding the mechanisms of protein recognition at the molecular level and to unravel the global picture of protein interaction in the cell. Protein interactions are fundamentally characterized as stable or transient, and both types of interactions can be either strong or weak. Stable interactions are those associated with proteins that are multi-subunit complexes, and the subunits of these complexes can be identical or different. Hemoglobin and the core RNA polymerase are examples of multi-subunit interactions that form stable complexes.

On the other hand, transient interactions are expected to control the majority of cellular processes. As the name implies, transient interactions are temporary in nature and typically require a set of conditions that promote the interaction, such as phosphorylation, conformational changes or localization to discrete areas of the cell. Transient interactions can be classified as strong or weak, and fast or slow. While in contact with their binding partners, transiently interacting proteins are involved in a wide range of cellular processes, including protein modification, transport, folding, signaling, and cell cycling. Therefore, a study of protein interaction networks is important not only from a theoretical stance but also in terms of potential practical applications, because it might enable new drugs to be developed that can specifically interrupt or modulate protein interactions.

Many experimental methods have been developed for identification of protein interactions. Some of the experimental methods enable screening of a large number of proteins in a cell. Such methods include yeast two-hybrid (Y2H), tandem affinity purification (TAP), mass spectroscopy (MS), DNA and protein microarrays, synthetic lethality, and phage display. Other methods focus on monitoring and characterizing specific biochemical and physiochemical properties of a protein complex. Despite this, a complete interaction network for many organisms is not available. The low interaction coverage along with the experimental biases toward certain protein types and cellular localizations reported by most experimental techniques, call for the development of computational methods to predict

whether two proteins interact. This include methods based on
(i) the co-localization of potentially interacting genes in the same gene clusters or protein chains (gene cluster, gene neighborhood, and Rosetta stone methods),
(ii) co-evolution patterns in interacting proteins (sequence co-evolution methods), and
(iii) the co-expression of genes. Some methods find patterns of co-occurrences in interacting proteins, protein domains, and phenotypes (phylogenetic profiles and synthetic lethality methods), while others use the presence of sequence/structural motifs characteristic only for interacting proteins (classification methods, association methods). These methods can be very useful for choosing potential targets for experimental screening or for validating experimental data and can provide information about interaction details, in the case of domain prediction methods which might not be apparent from the experimental techniques. However, these methods may not be generally applicable to all proteins in all organisms, and may also be prone to systematic error. Recently, a number of complementary computational approaches have been developed for the large-scale prediction of protein-protein interactions based on protein sequence, structure and evolutionary relationships in complete genomes. In the following sections, we report on experimental approaches used for identification of protein-protein interaction, in addition to complementary computational methods for PPI predictions.

1.1 Compartmentalization of PPI

Proteinprotein interaction data are one of the most valuable sources for proteome-wide analysis , especially to understand human diseases on the systems-level (Vidal et al, 2011) and to help network-related drug design (Bulusu et al, 2014). However, These protein interactomes in particular have been susceptible to questions about their biological meaning. Interaction data often contain interactions, where the two interacting proteins have no common subcellular localizations (Wiwatwattana and Kumar, 2005). These interactions could be biophysically possible, but biologically unlikely (Levy et al, 2009). Thus, these interactions cause data bias that leads to deteriorated reliability in interactome-

based studies, especially those involving subcellular localization-specific cellular processes (Lee et al, 2014).

1.2 Experimental Methods for Detecting PPIs

Various experimental methods for detecting protein-protein interactions have been developed. This include techniques that enable screening of a large number of proteins in a cell, such as yeast two-hybrid (Y2H) (Fields and Song, 1989), tandem affinity purification (TAP) (Krogan et al., 2006), mass spectroscopy (MS), DNA and protein microarrays, synthetic lethality (Ooi et al., 2006), and phage display (Mullaney and Pallavicini, 2001). Other methods focus on monitoring and characterizing specific biochemical and physiochemical properties of a protein complex (Benjamin et al., 2007).

1.2.1 Yeast Two-Hybrid Assays

Yeast two-hybrid is based on the reconstitution of a functional transcription factor when two proteins or polypeptides of interest interact. This takes place in genetically modified yeast strains, in which the transcription of a reporter gene leads to a specific phenotype, usually growth on a selective medium or change in the color of the yeast colonies. The Y2H method is based on the fact that many eukaryotic transcription activators have two different domains namely the DNA binding domain (DBD) that recognizes a specific DNA sequence, and the activation domain (AD). The AD coordinates the assembly of the elements required for transcription and enables RNA polymerase II to transcribe a specific reporter gene downstream of the DBD domain. The transcription is inactivated by splitting the DBD and AD, but transcription can be restored if a DNA-binding domain physically interact with an activating domain (Benjamin et al., 2007). Using the yeast two-hybrid system the protein of interest (X) is expressed as a fusion protein to the DBD (DBD-X; also known as the bait protein) and the activation domain is fused to the second protein of interest (Y), (AD-Y; also known as the prey protein). The AD-Y fusion vector is introduced into a yeast strain containing the DBD-X fusion partner by transformation

or mating. Only if proteins X and Y physically interact with one another are the DBD and AD brought together to activate expression of the downstream reporter gene (Figure 1).

1.2.2 Affinity Purification-Mass Spectrometry

Affinity purification-mass Spectrometry (AP-MS) is a powerful method of studying novel interactions. AP-MS experiments allow identification of PPIs in a complex. The method involves biochemical isolation of protein complexes using an inherent interaction and subsequent identification of their constituting proteins using mass Spectrometry (Kimet al., 2010). In a AP-MS experiment, a protein of interest (bait) is firstly tagged and expressed in *vivo*, then followed by the Immunoprecipitation of the bait plus its interacting partner (preys). Lastly the preys are identified using mass spectrometry based on their mass-to-charge ratios. The main challenge in AP-MS is identification of the real interactors from the many positive bait prey combinations. An advantage of these methods is that several members of a complex can be tagged simultaneously. However these techniques may miss complexes that are not present under certain conditions.

1.2.3 Synthetic Lethal Screens

Synthetic lethality describes any combination of two separately non-lethal mutations that leads to the non-availability of an organism, (Ooi et al., 2006). Ordinarily, individual mutations are compensated for or buffered. Synthetic lethal relationships can occur in genes acting in the same biochemical pathway or those in linked pathways (Tong et al., 2001; Tong and Boone, 2006). As such, synthetic screens detect functional linkages between two proteins. The technique's applicability in high throughput PPIs mapping is hampered by its complexity.

1.2.4 Phage display

This method is based on the ability of bacteriophage to present engineered proteins on their surface coat (Mullaney and Pallavicini, 2001). Interactions between phage-displayed proteins and target proteins can be rapidly identified and characterized using high throughput methodologies that makes the method attractive for scanning large peptide libraries. By generating a phage library, a large set of proteins can be screened for interactions.

1.2.5 Protein microarray

Protein microarray is steadily gaining popularity for PPI investigation. Protein microarray consists of proteins that are immobilized in a grid-like pattern on small surfaces (MacBeath and Schreiber, 2000). Huge numbers of proteins are used to screen and assess patterns of interaction with samples containing distinct proteins or classes of proteins. Two classes of protein microarrays are currently available: analytical and functional protein microarrays. In analytical protein microarrays, well-characterized molecules with specific activity, such as antibodies are used as immobilized probes. On the other hand, functional protein microarrays are mainly applied in areas of biological discovery such as drug target identification and validation, protein interaction and immune responses. For example (Zhu et al., 2001) cloned 5800 open reading frames on yeast proteome using protein microarray method. Therefore, they identified a number of new calmodulin and phospholipid interacting proteins. A common potential binding motif was identified for many of the calmodulin-binding proteins. However, this method has some drawbacks including (i) finding a surface and a method of attachment that allows the proteins to maintain their secondary or tertiary structure and thus their biological activity and their interactions with other molecules, (ii) producing an array with a long shelf life so that the proteins on the chip do not denature over a short time, and (iii) identifying and isolating antibodies or other capture molecules against every protein in the human genome.

1.3 Computational Methods for Predicting PPIs

The labor intensive experimental techniques for the detection of PPIs (see section 1.1) may not be generally applicable due to time constraints and high cost of experiments. Recently, a number of computational methods have been developed for the large scale prediction of PPIs based on protein sequence, structure and evolutionary relationships in complete genomes. In this section, we will describe computational methods and resources available for protein-protein interaction prediction that exploit the structural, genomics and biological contexts of proteins in complete genomes. In addition to algorithms and methods for interaction prediction, a number of useful databases pertaining to protein-protein interactions will be described. These databases combine a large amount of data from both computational and experimental techniques.

1.3.1 Methods Based on Single Biological Evidence

1.3.1.1 Gene Neighbor

One of the first methods of predicting PPIs using single biological evidence is gene neighbor or co-localization. This method utilizes the idea that genes which physically interact will be kept in close physical proximity to each other within the genome (Tamames et al., 1997; Overbeek et al., 1999; Skrabanek et al., 2008). For example, in prokaryotes, related genes are often co-localized into regions called operons. Genes involved in the same biological process or pathway are frequently situated in close proximity. Hence it is possible to predict physical interaction between genes that are in close proximity (e.g. 500bp). Many studies have been conducted for PPI prediction using this method. For example, Okuda et al. (2005) examined the conservation of gene co-regulated between two distantly related prokaryotes, *Bacillus subtilis*(B.subtilis) and *Escherichia coli* (E.coli). The analysis shows that about 60-80 % of gene pairs conserved co regulation relationships. In addition, pathway and Clusters of Orthologous Groups (COG) analyses demonstrated that conserved co-regulated gene pairs share the same functions.

The study by Tamames et al. (1997) analyzed genomes of *Haemophilus* influenza and

Escherichia coli to study gene order relationship and genome organization. Their study showed that functionally related genes are often transcribed as a single unit, an operon, in bacteria and are co-regulated in eukaryotes. In addition Dandekar et al. (1998) applied a systematic comparison of nine bacterial and archaeal genomes. Their study showed that the proteins encoded by conserved gene pairs appear to physically interact in bacterial and archaeal genomes.

1.3.1.2 Phylogenetic Profile Methods

Another powerful form of single biological evidence is the phylogenetic profile method, which is based on the hypothesis that, non-homologous interacting, and functionally linked proteins co-evolve and have orthologs in the same subset of fully sequenced organisms (Pellegrini et al., 1999; Benjamin et al., 2007). A phylogenetic profile for each protein is constructed based on the presence or absence of that protein across a range of genomes. The presence/absence of a given protein in a given genome is indicated as 1 or 0 respectively at each position of a profile (Figure 1.1). Proteins or their profiles can then be clustered using a bit-distance measure, and those proteins from the same cluster are considered potential interacting partners, (Benjamin et al., 2007). The disadvantage of this methodology is that it fails to correctly classify ubiquitous proteins, i.e, proteins that are present in all genomes but are not necessarily functionally linked. Additionally, evolutionary processes such as gene duplication, loss, and horizontal gene transfer could hamper accurate construction of phylogenetic profiles.

[here] **Schematic diagram of phylogenetic profiling**

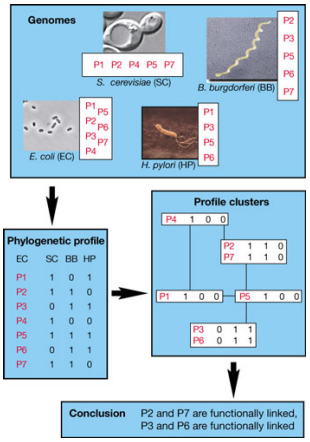


Figure 1.1: **Schematic diagram of phylogenetic profiling**: For each E. coli protein, the profile is constructed, indicating which genomes code for homologs of the protein. Profiles are calculated to determine which proteins share the same profiles. Proteins with identical (or similar) profiles are boxed to indicate that they are likely to be functionally linked. Boxes connected by lines have phylogenetic profiles that differ by one bit and are termed neighbors. The figure was adopted from the URL($http$: $//www.nature.com/nature/journal/v405/n6788/box/405823a0_bx1.html$)

1.3.1.3 Gene fusion

Single biological evidence methods for predicting PPIs also include the analysis of gene fusion across complete genomes. Gene fusion or Rosetta stone (Figure 1.2) is a hybrid gene formed from two separate genes, and can occur as a result of translocation, interstitial deletion, or chromosomal inversion. This method is complementary to both co-localization of genes and phylogenetic profiles, and uses both gene location and phylogenetic analysis to infer function or interaction (Enright et al.,1999; Skrabanek et al., 2008). The gene fusion approach predict PPIs from different genomes (Benjamin et al., 2007) based on the principle that interacting proteins/domains have homology in other genomes that are fused into one protein chain. Gene fusion events are detected by multiple species sequence comparisons. This method starts by searching for unfused protein sequences that are

Schematic diagram of gene fusion method

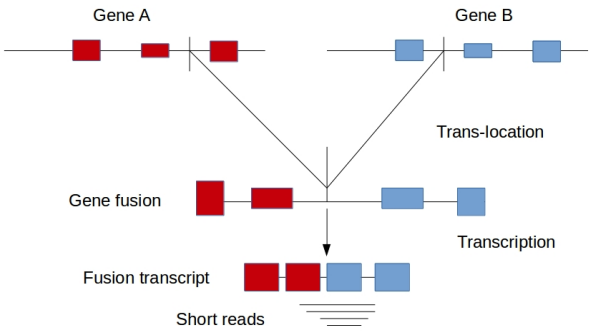


Figure 1.2: **Schematic diagram of gene fusion method**: Thus proteins in a genome are predicted to interact if they are fused together into a single protein (Rosetta protein) in another genome.

homologous to fused proteins in a given reference genome, but not to each other. Then, the resulting unfused protein sequences are aligned to different regions of the reference proteins, showing that the reference protein is a result of a gene fusion.
This method has been successfully applied to a large number of genomes (Enright et al., 2002; Marcotte et al., 1999). Enright et al. (1999) predicted functional associations of proteins using this method. Their analysis detected 215 genes or proteins in the complete genome of *E.coli* and *Haemophilus influenza*. In addition they predicted 39730 functional association pairs from 24 fully sequenced genomes using a gene fused approach.

1.3.1.4 Domain Profile Methods

Techniques in this class utilize conserved sequence properties such as domains, motifs and signatures to predict interactions. Thus a sequence signature is defined as a "highly conserved region", a sequence pattern that is found repeatedly in a group of related protein sequences [15]. By this definition, a sequence signature could be a protein family,

functional domain, functional site, or any conserved region of unknown function, and thus the actual physical manifestation of a signature can vary greatly in size. Sprinzak and Margalit, (2001) used sequence signatures found in experimentally determined interacting protein pairs in yeast to predict PPIs. Deng et al. (2002) developed an optimization method termed maximum likelihood estimation (MLE) that infer domain interactions by maximizing the likelihood of the observed protein interaction data. The expectation-maximization (EM) algorithm is used for the optimization of the probabilities of domain interactions. They apply this method to predict cellular functions (43 categories including a category 'other') for yeast proteins defined in the Yeast Proteome Database(YPD), using the protein-protein interaction data from the Munich Information Center for Protein Sequences (MIPS, http://mips.gsf.de).

1.3.2 Methods Based on Multiple Sources of Biological Evidence

In the previous section we explored the first group of genomic context methods for predicting PPIs. In order to improve the accuracy of prediction, the second set of methods utilize multiple biological evidences simultaneously. These sources individually are usually weakly associated with the interaction but can yield reliable predictions when analyzed as a group. Studies that utilize multiple evidence, formulate the PPI prediction as a binary classification problem and solve the task with a classifier. In this manner the classifier is trained to distinguish between positive sets that are truly interacting proteins, and negative sets or non-interacting pairs. Protein sequences of different lengths should be converted into feature vectors of the same length, where features refer to a particular information source regarding either protein interaction partner.

The advantage of classification is the ability to assess any feature's predictive power using feature selection, and it can handle missing data which is common in biological datasets. In order to perform binary classification using supervised learning approaches, the classifier requires positive and negative data sets for training, (Bishop 2006, Mitchell 1997).

One of the main challenges for binary classification of PPIs prediction is the selection of the negative data. In the absence of a gold standard negative data set that is experimentally validated, many studies suggest different methods for generating negative data. One of the common methods for choosing negative data for training a predictor of protein-protein interactions is based on annotations of cellular localization, and the observation that pairs of proteins that have different localization patterns are unlikely to interact, (Jansen et al., 2003), (Jansen and Gerstein, 2004). Some other studies selected non-interacting pairs uniformly at random from the set of all protein pairs that are not known to interact (Gomez et al., 2003; BenHur and Noble, 2005). These approaches are likely to yield their own biases (BenHur and Noble, 2006), but the uniformly random selection method is preferred. A number of classification techniques have been applied for predicting PPIs. Such methods include Naive Bayes, Random forest, Support Vector Machine and Artificial Neural Network (Zahiri et al., 2013).

1.3.3 Naive Bayes

Naive Bayes is a machine learning classifier that apply the Bayes theorem with an implementation that is computationally efficient and easy to interpret. In addition, the methods is ideal for problems that involve a normal distribution. Naive Bayes can be trained with small training data for supervised learning tasks using maximum likelihood, but it will not work well in complex problems (Najafabadi and Salavati, 2008; Zahiri et al., 2013). Many studies have used naive Bayes for PPIs such as (Jansen et al., 2003; Lu at al., 2005; Liu et al., 2012).

1.3.4 Random Forests

Random forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees (Breiman, 2001). Chen and Liu, (2005) developed

a random forest model for predicting PPIs on *Saccharomyces cerevisiae* dataset with sensitivity 79.78% and specificity 64.38%. This is a significant improvement over the (Breiman, 2001) prediction that used maximum likelihood approach (Dayer et al., 2008).

1.3.5 Support Vector Machine

Support vector machine (SVM) is a supervised learning model used for classification and regression analysis (Vapnik, 1995; Cristianini and Taylor, 2000). An SVM model maps the examples into space as points, see Figure 1.3, so that the examples of the separate categories are divided by a clear margin that is as maximum as possible. Unlabeled data are then mapped into that same space and predicted to belong to one of the two categories. SVM model is used in computational biology for classifying biological data, as well as protein-protein interaction predictions (Bock and Gough, 2001; Gomez et al., 2003). SVM is powerful and can classify problems with arbitrary complexity, but also require large memory. In addition, in order to build a good SVM model the hyper parameter must be optimized (Ben Hur et al., 2008). Many authors have used SVM for PPIs prediction. A study conducted by Bock and Gough (2001) developed and SVM model for PPI predictions using protein primary structure. They trained the model with physiochemical properties of amino acid, including charge, hydrophobicity, and surface tension for each residue in the sequence. Their model achieved an average 80% of accuracy. In addition, Bradford and Westhead (2005) applied SVM combined with surface patch analysis for protein-protein binding site prediction using data from the protein data bank database (PDB)(Berman et al,. 2000) for protein secondary structure. More than 75% accuracy was achieved after training the SVM with cross validation strategy. Gui et al (2008), made a further attempt to improve the performance of the SVM for PPI prediction by adding new feature sets which are physiochemical properties such as hydrophobicity, hydrophicility, volumes of side chains of amino acids, polarity, polarizability, solvent-accessible surface area (SASA) and net charge index (NCI) of side chains of amino acids. The method has been tested on *Saccharomyces cerevisiae* PPIs datasets which are

experimentally validated, and achieved an accuracy 88%.

SVM-Support Vector Machines

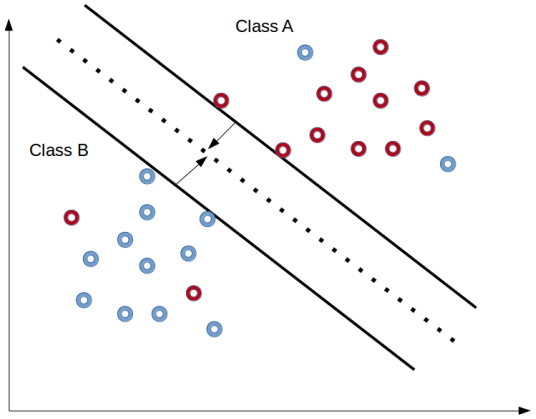


Figure 1.3: **Maximum-margin hyperplane and margins for an SVM trained with samples from two classes**: Samples on the margin are called the support vectors. The SVM learned the representation of a hyperplane, here illustrated through an enclosed rectangle that best separates the two classes of examples from each other. The examples that lie on the outside edge of the hyperplane are the so called support vectors (the actual representation learned by the SVM)

1.3.6 Neural Network

Artificial neural network (ANN) is a statistical learning algorithm inspired by biological neural networks. The idea of artificial neurons was found by McCulloch and Pitts, (1943) and developed by Werbos (1974). There are many types of ANN, but the most widely used is the multilayer feed forward neural network. An artificial neural network is a black box approach that has been used successfully in predictive modeling. Initially, the neural network must be trained. For the purpose of training, all the characters describing the unknown situation must be presented to the neural network, along with their predictions (also given). There are many types of neural network algorithms. In this study we used

the multi-layer feed-forward neural network (MFFN) (Figure 1.4). MFFN is used more frequently than other neural network types for a wide variety of classification and prediction tasks. A MFFN consists of neurons or nodes that are ordered into layers. The first layer is called the input layer, the last layer is called the output layer and the layers in-between are called hidden layers. An MFNN can have more than one hidden layer. Each layer in the MFFN is connected with other layers through weights which control the signal transfer between nodes through the so-called transfer or activation function. The training of an MFFN is toward searching for optimal values of the weights. For the activation function $f(x)$, the input I_k to node k is the weighted sum of the outputs of all nodes $(j = 1, 2, ..., n)$ connected to it. Here O_k is the output of the node k, w_{kj} is the linking weight between nodes k and j, and d_k is a bias.

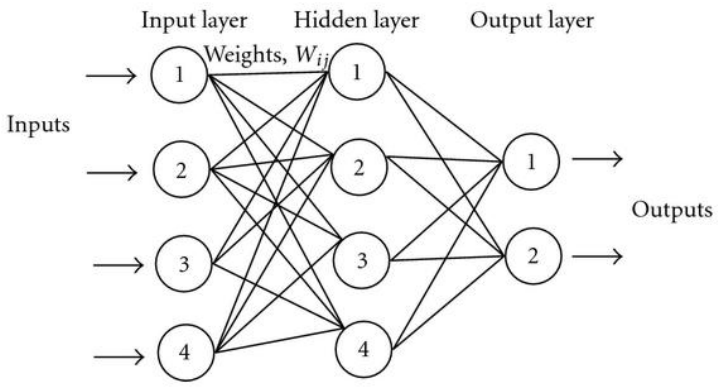


Figure 1.4: **An artificial neural network depicted as an interconnected group of nodes, akin to the vast network of neurons in a brain**: Here, each circular node represents an artificial neuron and an arrow represents a connection from the output of one neuron to the input of another.

1.3.7 Computational Methods for Inter-Species PPIs

The knowledge of host pathogen PPIs is crucial for understanding the pathogenesis of the relevant disease (Bosch et al., 1998; Mogensen et al., 2006). However, experimental resources for studying interactions between host and pathogen proteins are rather limited.

Several computational methods for predicting interspecies PPIs have been developed, including methods based on interolog , interacting domain/motif, structure, and even machine learning (Zhou et al., 2013).

1.3.7.1 interolog Based Approach

Interolog based methods constitute the conventional way of predicting host-pathogen interactions. The methods are based on the hypothesis that pairs of interacting proteins in one species are expected to be conserved in related species. The rationale behind this approach is that if two proteins interact in one organism, their interolog in another organism have a higher chance of interacting. This is based on the assumption that sequence and structural similarities between gene products suggest functional similarities. Therefore, the interolog based method for host-pathogen PPIs prediction procedure proceeds as follows: (i) selecting a known pair of PPI (A, B) in some source species (the template PPIs), (ii) find in the host a interolog (A') and in the pathogen a interolog (B'), respectively, of the two (A, B) in the template PPIs and, (iii) predicting that (A', B') interact.
The interolog based approach has been applied to many studies, for example

Tyagi et al. (2009) developed interolog based methods to predict PPIs between human host and *Helicobacter pylori*. They identify 623 *Helicobacter* proteins that interact with 6559 human proteins. Further analysis shows that most of their predicted *Helicobacter pylori* proteins are known to be secreted proteins. Moreover, Krishnadev and Srinivasan, (2011) used a interolog based method to predict protein protein interaction between human and three pathogens namely (*E. coli, Salmonella enterica typhimurium* and *Yersinia pestis*). They implement the methods on data extracted from protein interaction database (DIP)(Ioannis et al., 2002) and a database of protein family and domain interactions found

in the Protein Data Bank (iPfam) (Finn et al., 2014). Thus they identify several host-pathogen protein interactions, most of which are disease related or hypothetical proteins. (Wuchty., 2011) attempt to improve the performance of interolog based methods by incorporating a machine learning approach coupled with expression and molecular characteristics. First they implement the approach on human *Plasmodium falciparum* PPI prediction. Thereafter, their analysis on the predicted sets shows that parasite proteins tend to target central proteins in order to take control of the human cell. In addition they target human proteins involved in pathway signaling and regulation.

1.3.8 Machine Learning Approaches

Machine learning techniques (supervised and semi supervised) have been applied intensively for interspecies PPI predictions. However, these methods require template PPI data sets associated with appropriate biological and biochemical properties as features for training and testing purposes. Many studies utilize this technique. For example, Tastan et al. (2009) developed a Random Forest classifier to predict PPIs between human and HIV-1 by incorporating multiple features sets such as interacting domain, gene ontology annotations, post-translation modifications, tissue distribution, gene expression, and topological properties of the human protein in the interaction network. The extension of their work using semi-supervised learning is presented by Qi et al. (2010). They incorporate direct interaction data with likely interactions, which are not experimentally validated to train a classifier for human HIV-1 PPIs prediction. However, they perform multi-task learning on supervised classification using labeled data (truly interaction) and semi-supervised task for partially labeled data (no evidence of direct interaction). Thus, they train a multi-layer perceptron network using labeled data for PPI prediction while the semi-supervised learning task share the network layer of supervised classifier. Therefore, the performance shows improvement compared to their previous work. Furthermore, Dyer et al. (2008) integrated known intra-species PPI data with protein-domains profiles to predict PPIs for human-*Plasmodium falciparum*, and used Bayesian statistics to assess

the prediction. Many studies have been done on human-virus interactions because of the abundance of high-throughput experimental data. A support vector machine (SVM) combined with linear kernel have been developed by Dyer et al. (2011) for PPI prediction between human and HIV. The authors explored different types of features including domain profile, protein sequence k-mers, and graph theoretic properties of human PPI network. Consequently, their model achieved a precision value of 70% and recall greater than 40% when they used a combination of all three features. Likewise, Cui et al. (2012) proposed an SVM, based on feature: three consecutive amino acid frequency of sequence feature to predict human-virus PPIs. Their study showed the importance of three consecutive features on model performance, which achieved above 80% accuracy. Qi et al. (2010) proposed a solution to the lack of training data by using semi-supervised learning for host-pathogen PPIs. They combined true positive data with partial positive (indirect interactions) as a training set. However, high rates of false positives are likely to increase when using partial sets. In the case of improving supervised learning performance with the lack of reliable training data, an interesting question to ask is whether we can extend the model by using a training set, where host-pathogen data originates from the same host but different pathogens.

1.3.9 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 those host-pathogen protein pairs that match some template PPI are interacting (Smith and Sternberg, 2002).

1.3.9.1 Comparative modeling

Prediction by comparative modeling is a representative structure-based approach. For example, in a study by Davis et al. (2006) an automated pipeline for large-scale comparative protein structure modeling, 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 34 super families that the protein belongs to are identified. A database of protein structural interfaces, PIBASE, is then scanned. If a SCOP super family of a host protein and a SCOP super family of a pathogen protein are both involved in the same PIBASE 35 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 above process. In this case, template interactions in alternative databases (e.g. IntAct) are considered by Davis et al. (2006) Specifcally, 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 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, which results in a significant reduction of potential host-pathogen PPIs by several order of magnitudes.

1.3.9.2 Structural similarity

Structural similarity can also be analyzed using the Dali database (Holm and Rosenstrm, 2010). This strategy has been adopted to predict human-HIV PPIs, human-*Dengue virus* (DENV) PPIs and *Aedes aegypti*-DENV PPIs. Dali calculates a structural similarity score by comparing the 3D structural coordinates of two PDB entries. To predict the human-HIV and human DENV PPIs, structurally similar pathogen (HIV, DENV) and host human proteins are determined using the structural similarity method. Then, under the assumption that pathogen proteins having similar structure to host proteins are likely

to participate in the similar set of PPIs (human PPIs dataset from HPRD31), the pathogen proteins are directly mapped to their high-similarity matches within the host intra-species PPIs network in order to predict the host-pathogen PPIs. The same structural similarity prediction method has been applied to identify orthologs between *Drosophila melanogaster* and *Aedes aegypti* and map *D. melanogaster* DENV PPIs to predict. The accuracy of this prediction method depends on the performance of Dali in determining structurally similar pathogen and host proteins. The available information on pathogen and host protein structures and the quality of host intra-species PPIs data also have a significant influence on prediction results.

1.4 Protein interaction data resources

The rapid accumulation of PPIs data has necessitated the development of advanced storage systems. Although a plethora of repositories serving this purpose have been developed, a few select databases provide consistent, reliable interaction data. These databases can be grouped into three classes: (i) those have that store manually curated PPIs, (ii) those that store predicted PPIs, and (iii) those that store both curated and predicted PPIs.

Table 1.1: Databases that store PPI data

Database	Description
DIP	Catalogs experimentally determined interaction between Proteins by combining information from different sources http://dip.mbi.ucla.edu/dip/
IntAct	The data available in the database originated entirely from published literature and it is manually curated http://www.ebi.ac.uk/intact/.
MINT	MINT focus exclusively on curation of physical protein interactions and ignores computationally predicted interactions http://mint.bio.uniroma2.it/.

BioGrid	It is a database of physical and genetic interactions, that incorporates both high throughput data and curated data https://thebiogrid.org/.
STRING	Search tools for the retrieval of interacting gene/proteins nd it contain both known and predicted data http://string-db.org/.
BIND	Archives three type of interactions namely molecular interaction, complexes, and pathways http://bind.ca.
VirusMINT	Collecting all interactions between viral and human proteins, reported in the literature http://mint.bio.uniroma2.it/virusmint/.

1.5 Problem formulation

Despite decades of drug research and development, infectious diseases are still resulting in millions of deaths each year. Research efforts are ongoing for better understanding of the mechanism by which the pathogen invade the host cell, and identification of potential drug targets. Therefore, protein-protein interaction form the foundation of communication between a host and a pathogen and play a major role in infection (Dyer et al., 2008). However, identification of physical interaction between proteins was limited to labor-intensive experimental techniques such as co-precipitation or affinity chromatography (Benjamin at al., 2007). Several experimental assays that probe interactions in a high-throughput manner are now available and such methods include the yeast two-hybrid screen and methods based on mass spectrometry. These methods however, may not be generally applicable to all proteins in all organisms, and may also be prone to systematic error. Recently, a number of computational methods have been developed for the prediction of protein-protein interactions based on protein sequence, structure and evolutionary relationships among completely sequenced genomes. Furthermore, many computational methods have been developed for protein-protein interaction prediction within a single organism (intra-species) and across species (inter-species).

Some of the methods to predict intra-species PPI depend on genomic inference. Such methods utilize the idea that physically interacting genes will be kept in close proximity to each other on the genome (Bowers et al., 2004; Dandekar et al., 1998; Overbeek et al., 1999; Galperin et al.,2000). Co-localization of genes across multiple genomes is an indicator of physical interaction between the encoded proteins. The study by Tamames et al. (1997) analyzed genomes of *Haemophilus* influenza and *Escherichia coli* to study gene order relationship and genome organization. Their study shows that functionally related genes tend to be localized in close proximity, and different from unrelated genes. In addition, Dandekar et al. (1998) shows that the proteins encoded by conserved gene pairs appear to be physically interacting in bacterial and archaeal genomes. Another approach based on genomics context is phylogeny profiling (Galperin et al., 2000; Pellegrini et al, 1999; Snitkin et al,2006), which is based on comparison of evolutionary distance between the sequences of associated protein family (Pazos et al.,2001). A method that relies on the gene fusion event (Marcotte et al., 1999; Enright et al., 1999; Marcotte et al., 2002) was introduced following the observation that very often, pairs of interacting proteins have homology in another organism fused into a single protein chain (Enright et al., 1999; Huynen et al., 2000; Marcotte et al., 1999). It should be noted, that the above mentioned methods are not without flaws, and could result in an undesirably high rate of false positives. There are other computational methods that depend on three-dimensional aspects of protein interaction, but the scarcity of three-dimensional protein data makes the utility of these methods limited. In addition, different classification approaches have been applied to the prediction of protein-protein interactions. These methods use a variety of biological information items to train a classifier to distinguish between positive examples of truly interacting proteins pairs from the negative examples of non-interacting proteins. For example, Guo et al. (2008) developed a classifier using support vector machine combined with auto-covariance for the prediction of PPIs in *Saccharomyces cerevisiae* and the methods achieved accuracy of 88.09%. Yet, other studies utilize supervised classification techniques, such as as Random Forest (RF) (Chen and Liu, 2005), Neural Network (Fariselli et al., 2002; Eom and Zhang, 2005), SVM (Zhang et al., 2014; Shen

et al., 2007; You et al., 2004). On the other hand computational methods for predicting host-pathogen protein protein interaction has not received much attention due to the difficulty of experimental validation. Nevertheless, computational approaches such as Bayesian network and SVM have been used successfully to predict host-pathogen PPIs, (Dyer et al., 2007; Krishnadev et al., 2011) respectively. For example, Dyer et al. (2008) integrated known intra-species PPI data with protein-domain profiles to predict PPIs between human-*Plasmodium falciparum*, and used Bayesian statistics to assess the prediction. Tastan et al. (2009) used a random forest classifier to predict PPIs between human and HIV-1 by incorporating multiple features sets such as interacting domain, gene ontology annotations, post-translation modifications, tissue distribution, gene expression, and topological properties of the human protein in the interaction network. Another study by Wuchty, (2011), used a random forest classifier to predict PPIs between human and *Plasmodium falciparum* where researchers validated the results using co-expression data of human genes in the presence of parasites. Computational approaches have been applied to human-virus interactions because of the abundance of high-throughput experimental data. We have also mentioned the paper by Qi et al. (2006), using semi-supervised learning for host-pathogen PPI prediction, based on true positive data as well as partial positive data. However, one can expect high rates of false positives due to the partial positive sets. In summary, previous studies differed in terms of classifiers, feature sets, and their encodings and gold-standard datasets used. Thus, in order to develop a high accuracy classification model, careful consideration must be taken in the selection of the classifier, feature sets, and training data. Despite the vast amount of genomic data available today, there is a lack of experimentally validated host-pathogen PPI data for most pathogenic organisms , specifically human-bacteria interactions. Therefore, the challenge remains for studying host-pathogen protein-protein interaction for human-*Mycobacterium tuberculosis*, where experimentally verified PPIs data is less than 200 pairs. Nevertheless, in this study a sequence based prediction model is proposed by using Artificial Neural Network (ANN), with a novel combination of features generated from literature to enhance the accuracy of the model results. The challenge is formidable when it comes to

the study of host-pathogen protein-protein interaction for human-*M-tuberculosis* PPI data set contains less than 200 experimentally verified pairs. To overcome this challenge, we suggest a computational predictor model of the human-*Bacillus anthracis* PPIs, in which case, sufficient experimentally validated PPI data is available. We set up a computational predictor of host pathogen PPIs for the human-*Bacillus anthracis* case. In the latter case, sufficient experimentally validated PPI data is available. Then for the same human-*Bacillus anthracis* case, we test a different combinations of feature sets. The model model with quadruple featute sets is found to perform quite well, very much in step with the original prediction, but is less reliant on experimentally validated host-pathogen PPIs. This alternative model is then harnessed to predict human-*Mycobacterium tuberculosis* PPIs, and it is not very badly hampered by the shortage of experimentally validated human-*Mycobacterium tuberculosis* PPI data.

1.6 Research objectives

The objectives of this study were:

(i) Develop a binary classifier with an appropriate feature set for host-pathogen protein-protein interaction prediction using published human-*Hepatitis C virus* (HCV) PPI, and test the model on available host-pathogen data for human-*Bacillus anthracis* PPI.

(ii)Prediction of human-*Mycobacterium tuberculosis* PPIs using the feature set derived in (i) and a neural network.

(iii) Development of a web server for the prediction of human-*M-tuberculosis* PPI.

Chapter 2

Feature Optimization and its Application on Human-*Bacillus* PPI Prediction

2.1 Abstract

Background: Machine learning approaches have been successfully applied to the prediction of protein-protein interactions (PPIs) within a single organism i.e., intra-species PPI predictions. However, fewer studies have successfully applied these techniques to host-pathogen PPI, i.e., inter-species PPI prediction, due to limited experimentally validated PPI data for training and testing. These inter-species comparisons have focused primarily on human-virus interactions using different machine learning techniques and statistical models. Yet the selection of appropriate machine learning techniques and the choice of feature sets are important in order to strengthen the use of other biological datasets. **Results**: In this study quadruplet amino acids in combination with human interactome properties including graph-theoretic properties such as degree, cluster coefficient and betweenness centrality, and sequence similarity resulted in improved performance when an SVM classifier was used on a published human-HIV dataset. The same feature set was

used to assess machine learning approaches in predicting human-bacterial protein- protein interactions. The accuracy of the SVM approach that was applied to the human-HIV dataset was compared with a neural network approach to predict human-*Bacillus anthracis* protein-protein interactions. Our predictor shows an average accuracy of 93.4% when using quadruple and the human interactome features coupled with sequence similarity. The increased overall performance of our PPI prediction model using quadruple in association with network features, compared to triplets used in human-virus interaction, provides a refined set of target candidates. In summary, the results indicate that training neural networks with appropriate features, can improve host-pathogen PPI predictions. This algorithm was implemented using the neural network tool box of Matlab. Python scripts were used to extract features.

2.2 Overview

An important aspect of any host-pathogen system is the mechanism by which a pathogen infects its host. Host-pathogen protein-protein interactions play a vital role in initiating an infection. In particular, the proteins and molecules in cell surfaces form the foundation of communication between a host and a pathogen. PPIs constitute an important component of virtually every biological function on the molecular level. Consequently, unraveling the physical interaction between two proteins is essential for understanding the mechanisms of protein recognition at the molecular level and to understand the global picture of protein function in the cell. There are many experimental methods for detecting PPIs. They are expensive, labor intensive and time consuming. Experimental resources for studying interactions between host and pathogen proteins are rather limited. In view of such problems, computational methods for predicting PPIs can provide valuable complementary tools. Ranges of computational methods have been published that infer PPIs within single species (intra-species). A review of such literature can be found in Pitre et al. (2008). On the other hand, prediction of PPIs between hosts and pathogens (inter species) has not received the same attention. The knowledge of the protein interactions between host and

pathogen is crucial to understanding the pathogenesis of the relevant disease (Huang et al., 1998; Mogensen et al., 2006). Recently computational approaches were developed to infer PPIs between host and pathogen. For example, Dyer et al. (2008) integrated known intra-species PPIs data with protein-domain profiles to predict PPIs between human and Plasmodium falciparum. The application of machine learning techniques have been successfully applied to the prediction of human-virus interactions because of the abundance of high throughput experimental human-virus protein-protein interaction data sets. Qi et al. (2006) proposed a solution to the lack of training data by using semi-supervised learning for host-pathogen PPIs. They combined true positive data with partial positive data (indirect interactions) as training sets. However, high rates of false positives are likely when using partial sets. It is important to identify the features which are more relevant in computational prediction of interaction between a given pair of proteins. Not only does it help in revealing relationships between different data sources, but it can suggest which data should be generated by experiments to identify novel interactions in host-pathogen systems. Tastan et al. (2009) used a random forest classifier to predict PPIs between human and HIV-1 by incorporating multiple feature sets such as interacting domains, gene ontology annotations, post-translation modifications, tissue distribution, gene expression, and topological properties of the human interactome network. Another study, by Wuchty, (2011), used a random forest classifier to predict PPIs between human and Plasmodium falciparum where researchers validated the results using co-expression data of human genes in the presence of parasites. In order to develop a high accuracy classification model, consideration must be given to the selection of the classifier, features sets, and training data. With vast amounts of genomic data available today, there is a lack of experimentally validated host-pathogen PPI data for most model organisms, specifically human-bacteria interactions. Therefore, the challenge is formidable when it comes to the study of host-pathogen protein-protein interaction for where limited experimentally verified PPI data is available. For example, human *Mycobacterium tuberculosis* has less than 200 experimentally verified PPI pairs. To overcome this challenge we implement the following strategy. We set up a computational predictor of host pathogen

PPIs for the human-*Bacillus anthracis* case. In the latter case, sufficient experimentally validated PPI data is available. Then for the same human *Bacillus anthracis* case, we test a different features combination to predicting PPIs (at least in this particular case). The the quadruple features combination is found to perform quite well, very much in step with the original prediction, but is less reliant on experimentally validated host-pathogen PPIs. This alternative model is then harnessed to predict human-*Mycobacterium tuberculosis* PPIs, and it is not very badly hampered by the shortage of experimentally validated human-*Mycobacterium tuberculosis* PPI data. For the binary classification problem we use an artificial neural network, with a novel combination of features generated to enhance the accuracy of the predictor model. In addition, we assess the model performance, comparing the quadruple amino acid features combined with network features and sequence similarity that is utilized in this thesis against the triple amino acid features in existing literature. The triple features set in the literature were implemented using SVM. Therefore we test the curation between two feaure set using the published data set and the same algorithm (SVM).

2.3 Implementation

Prediction of protein-protein interactions using a supervised classifier requires training data. In the process of predicting PPIs, pairs of proteins are classified into two classes that can be labeled as interacting (positive) or not interacting (negative). The aim of the training step is to derive a representative sample of the spectral signatures for each class. The quality of the training data and the feature set significantly influence the performance of the algorithm that is being implemented using matlab, and this has an impact on the classification accuracy (DongMel and Douglas, 2002). The Matlab neural network toolbox provides algorithms implementation, functions, and apps to create, train, visualize, and simulate neural networks. Figure 2.1 shows the work flow for building the classification model and the feature selection process.

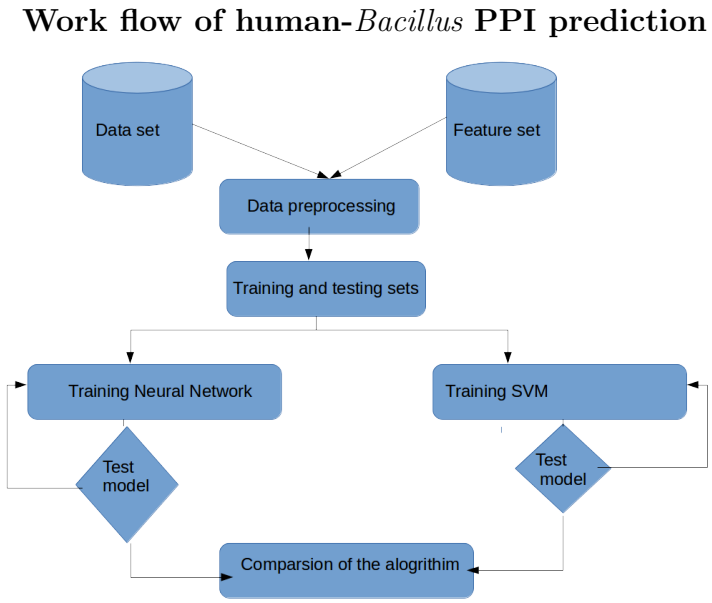


Figure 2.1: **Work flow of human-*Bacillus* PPI prediction**: The work flow starts with extracting host-pathogen PPI data, then converting the data to feature sets. Thereafter, building a model for PPI prediction.

2.3.1 Feature representation

Recent work by (Cui et al., 2012; Gomez et al., 2003; Taylor et al,. 2004) emphasized the value of encoding the important information content of the protein sequence for protein-protein interaction prediction. In addition, the protein sequences are of different lengths and should be converted into feature vectors of the same length, and the features from each protein are concatenated to form a single feature. The concatenation is made in this order, human protein feature + *M-tuberculosis* feature. In this study we consider four types of features namely consecutive amino acid triples, consecutive amino acid quadruple, sequences similarity and human interactome graph properties derived from the human interactome network.

2.3.1.1 Consecutive amino acid triples and quadruples.

The consecutive amino acid triples are the short amino acid sub-sequences of length 3 that occur in interacting proteins. The cardinality of the set of feature vectors, is approximately 8000 which is all possible triples combination from 20 amino acid. To reduce this high dimension, the 20 amino acids alphabet is reduced to six categories of biochemical similarity [IVLM, FYW, HKR, DE, QNTP, and ACGS] (Cui et al., 2012; Gomez et al., 2003; Taylor et al,. 2004). With this classification of amino acids, there are 216 possible amino acid triples from the above 6 groups. On the other hand there are 1296 possible sub-strings of length 4 using the 6 amino acid categories reported above. For both triples and quadruple we use a binary space (V, F) to represent proteins sequence, where V is a set of feature vectors with a fixed length (number of features) and F is a set of frequency vectors, the relative frequency value in a wider range makes it easier to discriminate protein sequences. A protein is first mapped to a feature vector v of fixed length. Then the feature vector v is mapped to a relative frequency vector q, the coordinates of which are defined by equation (2.1).

$$q_i = \frac{f_i - \min(f_1, f_2, ..., f_{216})}{1 + \max(f_1, f_2, ..., f_{216})} \tag{2.1}$$

Here f_i is the frequency of the i^{th} triple (respectively, quadruple) in the sequence, for $i = 1, 2, ..., 216$ (resp., $i = 1, 2, ..., 1296$).

2.3.1.2 Consecutive amino acid triples.

The consecutive amino acid triples are the short amino acid sub-sequences of length 3 that occur in interacting proteins. The cardinality of the set of feature vectors, is approximately 8000. To reduce this high dimension, the 20 amino acids alphabet is reduced to six categories of biochemical similarity [IVLM, FYW, HKR, DE, QNTP, and ACGS] (Cui et al., 2012; Gomez et al., 2003; Taylor et al,. 2004). With this classification of amino acids, there are 216 possible amino acid triples.

2.3.1.3 Sequence similarity.

For each pair of human-pathogen proteins, we use Emboss WaterCommandline to calculate a pairwise sequence similarity score. Therefore, we write a python script to filter the similarity score from the output file.

2.3.1.4 Human interactome graph properties.

Three graph property features were derived from topological properties of the human intra-species PPI network. These are the degree, clustering coefficient and betweenness centrality. The degree of a node in a network is the number of neighbors that are connected to it. Clustering coefficient is the ratio of the edges present among its neighbors to all possible edges that could be present between them. Betweenness centrality for a node is calculated as the fraction of shortest paths between node pairs that pass through the node of interest. In order to calculate the values of those properties, human interaction network data consisting of values of the mentioned properties for each protein was extracted (ftp://ftp.sanbi.ac.za/machine learning). A python script was written to map network properties to the training and testing data that was used in this study.

2.3.2 Neural network

In the previous section the new feature sets has better performance using neural network than SVM. The successful demonstration of the utility of our new feature set on human-HIV protein-protein interactions had to be tested on a bacterial system. The absence of sufficient experimentally verified huma-Mycobacterial protein-protein interaction data led us to consider a different bacterial pathogen namely *Bacillus anthracis*. We used a multi-layer feed-forward neural network (MFFN) combined with feature selection process to predict the protein interactions between human and *Bacillus anthracis* see section 2.4. An artificial neural network is a black box approach that has been used successfully in predictive modeling (Bishop, 2006). For the purpose of the initial step of training, all the characters describing the unknown situation must be presented to the neural network,

along with their classes label. There are many types of neural network algorithms. In this study we used the multi-layer feed-forward neural network (MFFN). The MFFN is used more frequently than other neural network types for a wide variety of classification and prediction tasks. A MFFN consists of neurons or nodes that are ordered into layers. The first layer is called the input layer, the last layer is called the output layer and the layers in-between are called hidden layers. Each layer in the MFFN is connected with other layers through weights that control the signal transfer between nodes through the so-called transfer or activation function. The training of an MFFN is to search for optimal values of the weights. For the activation function $g(x)$, the input I_k to node k is the weighted sum of the outputs of all nodes ($j = 1, 2, ..., n$) connected to it.

$$I_k = d_k + \sum_j W_{kj} O_j \tag{2.2}$$

$$O_j = g(I_j) \tag{2.3}$$

O_k is the output of the node k, W_{kj} is the linking weight between nodes k and j, and d_k is a bias.

Activation function.

$$Sigmoid(x) = \frac{1}{(1 + exp(-x))} \tag{2.4}$$

2.3.3 Support Vector Machine

We used a SVM to predict protein-protein interactions using the same dataset as published by (Cui et al., 2012; Gomez et al., 2003; Taylor et al,. 2004). We followed the same performance evaluation criteria used by Cui et al (2012) to evaluate the choice of triple versus quadruplet amino acids as par to the feature set. The criteria included sensitivity, specificity and accuracy. The sensitivity, also called the true positive rate or the recall rate, measures the proportion of actual positives that are correctly identified as such (e.g., the percentage of sick people who are correctly identified as having the condition), and in the set of actual positives, the subset of true positives is complementary to the false negatives. The specificity, sometimes called the true negative rate, measures the proportion

of negatives that are correctly identified as such (e.g., the percentage of healthy people who are correctly identified as not having the condition), and is complementary to the false positive rate. The accuracy of a measurement system is the degree of closeness of measurements of a quantity's to that quantitys actual (true) value. In machine learning, support vector machines (SVMs) (Cortes and Vapnik., 1995) are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. Given a set of training examples, each marked for belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non- probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they sit. In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces. When data is not labeled, a supervised learning is not possible, and an unsupervised learning is required, that would find natural clustering of the data to groups, and map new data to these formed groups. The clustering algorithm that provides an improvement to the support vector machines is called support vector clustering (Ben-Hur et al., 2001). SVM is highly used in industrial applications either when data is not labeled or when only some data is labeled as a preprocessing for a classification pass. In this thesis, a radial basis function kernel (rbf-kernel) is employed and is defined as

$$K(x_i, x_j) = exp(\frac{|x_i - x_j|^2}{\sigma^2}) \tag{2.5}$$

where x_i and x_j are the two vectors where one of them is a support vector and σ is an adjustable parameter that determine the area of influence of the support vector over the data space. Larger value of reduce the number of support vectors, since each support vector covers more data space. The SVM implementation used in the present study is SVMlight (Joachims, 1999). This program is freely downloadable from

http://svmlight.joachim.org/ . SVMlight has several hyperparameters which should be optimized in order to obtain a generative model.

2.3.4 Sub-network analysis of human-*Bacillus* interactions

Sub-networks of human-*Bacillus* proteins were generated using network analysis blog in within cytoscape software. The GO enrichment analysis was also done using cytoscape blogin namely Bingo (Shannon et al., 2003).

2.3.5 Performance evaluation

The receiver operating characteristic, or simply ROC curve, is a graphical plot which illustrates the performance of a binary classifier system. It is created by plotting the fraction of true positives out of the total actual positives (TPR = true positive rate) vs. the fraction of false positives out of the total actual negatives (FPR = false positive rate), at various threshold settings. A perfect test with 100% sensitivity and 100% specificity would show the curve tending towards the upper left corner.

We used ROC curves for both evaluating the performance of the feature selection and the prediction of PPI for human and *Bacillus anthracis*. Thus, we use human-*Bacillus* PPI data as a positive set and the negative sets were randomly generated as specified in section 2.4.2.

To evaluate the performance of our classification model we plotted the ROC curve using the ROCR R-package. In addition to the ROC curve we also use the *accuracy* to measure the model performance. The accuracy is the percentage of predictions that are correct.

We divide the training data into three sets: 60% for training, 20% for validation and 20% for testing. The first subset is the training set, used for computing the gradient and updating the network weights and biases. The second subset is the validation set. When the validation error increases for a specified number of iterations, the training is stopped, and the weights and biases at the minimum of the validation error are returned. The third subset is the test set. It is used to verify the network design.

2.4 Results and Discussion

2.4.1 Quadruplets contribute to improved feature selection

In this work we establish an optimal feature set by benchmarking consecutive quadruple amino acids, in combination with network features obtained from known human interaction graphs and sequence similarity, and compared these to triple amino acid features as reported by (Cui et al., 2012; Gomez et al., 2003; Taylor et al,. 2004). The comparison was standardized by using the training and testing data sets as used by (Cui et al., 2012; Gomez et al., 2003; Taylor et al,. 2004) and an SVM light classifier. We followed the same performance evaluation criteria used by (Cui et al., 2012; Gomez et al., 2003; Taylor et al,. 2004) to evaluate their model namely sensitivity, specificity and accuracy. The comparison results as reported in Table 2.1 shows that our model outperforms the model used by (Cui et al., 2012; Gomez et al., 2003; Taylor et al,. 2004) by 95.9 % to 80.5% in terms of sensitivity, 91.6% to 89.7% in term of specificity and 88.6% to 85.1 in term of accuracy. This performance shows the importance of the quadruple feature representation when combined with sequence similarity and human interactome network graph properties such as degree, betweenness centrality and cluster coefficient for improving host-pathogen PPIs prediction using the SVM classifier .

Table 2.1: **Performance of the model generated using triple feature (Cui et al., 2012; Gomez et al., 2003; Taylor et al,. 2004) in comparison with quadruple feature.**

	$^{e}SN(\%)$	$^{d}SP(\%)$	$^{c}AC(\%)$
Quadruplet feature set^{a}	95.9	91.6	88.6
Triplet feature set^{b}	80.5	89.7	85.1

$^{a}Features$ included consecutive quadruple amino acids.

$^{b}Features$ published by (Cui et al., 2012; Gomez et al., 2003; Taylor et al,. 2004) that

includes consecutive triple amino acids.
[e]*Sensitivity* (SN), [d]*specificity* (SP) and [c]*accuracy* (AC).

2.4.2 Model building and feature selection process

The ultimate goal of predicting human-*Mycobacterium tuberculosis* protein interactions is prohibited by the limited amount of experimentally verified human-*M-tuberculosis* protein protein interaction data. A collective PPI dataset for the human and the pathogen *Bacillus anthracis* provided a larger dataset to assess the accuracy of the newly established feature set (Table 2.1) on a bacterial pathogen instead of viruses. We used 554 human-*Bacillus anthracis* experimentally verified interacting pairs from the IntAct database (Henning et al., 2004). This data set served as a positive set for training the classifier. There is no gold standard negative set available for training and testing purposes. However, it is standard practice to create a negative dataset by choosing protein pairs randomly from the set of protein pairs that are not known to interact (Dyer et al., 2008; Tastan et al., 2011; Cui et al., 2012). The number of truly interacting pairs of human-*Bacillus anthracis* proteins is likely to be far less than the total set of proteins. These randomly generated protein pairs were filtered to ensure that there were no protein pairs that are known to interact in the positive dataset, python script can be found in ftp://ftp.sanbi.ac.za/machine-learning.

2.4.3 Performance of Triple amino acids in combination with network and sequence similarity features

The Matlab neural network toolbox was used to predicting human-*Bacillus* PPIs. The input data was randomly divided into three sets: 60% is used for training, 20% validation and 20% testing. The protein sequences were converted into a numerical feature representation that concatenate the triple feature with sequence similarity, and the three human interactome features. The model was trained using these features Table (2.2). The result

shows the performance of the triple feature and the combinations of triples with each of the other features in order to evaluate the importance of each single feature combined with triples. Table 2.2 shows the accuracies of the different feature combinations. The column labeled as model average shows combined average accuracy of the training, testing and validation, while the other columns present the training accuracy. From Table 2.2 we observe that the model average improves from 84.0% when using triple feature alone, to 91.3% when combining the triple feature with all other features. This result shows the importance of graph properties features together with the sequence similarity. Figure 2.2 visualizes the results presented in Table 2.2, using the ROC curve. As mentioned before, the more accurate the prediction the more the ROC curve will tend towards the upper left corner. In this case Figure 2.2 shows that the combination of triples with all other features performs best.

Table 2.2: Model performance (average accuracy) of the triple amino acid feature combined with different network features

Features	Model average	Training	Resting	validation
$three$[1]	84.0	87.6	79.9	77.9
$threeD$[2]	58.1	87.2	82.4	78.4
$threeB$[3]	83.7	87.5	75.0	74.5
$threeC$[4]	80.1	82.6	73.5	75.0
$threeS$[5]	83.4	85.4	78.4	78.9
$threeA$[6]	91.3	95.8	83.3	77.7

[1]*triple* consecutive amino acid frequencies.

[2]*triples+* betweenness centrality network properties.

[3]*triples+* clustering coefficient.

[4]*triples+* degree (network properties).

[5]*triples+* sequence similarity feature.

[6]*combination* of all the above-mentioned features.

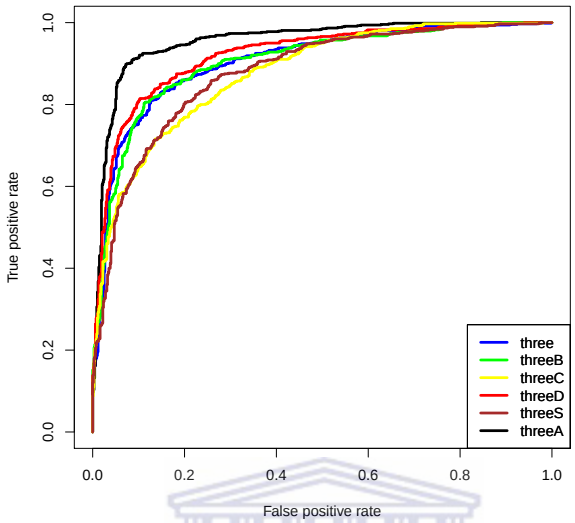


Figure 2.2: ROC curve for triple amino acids with other feature combinations

ROC curves for six different combinations of features. *Each curve indicates the feature set used. 'three' = triples consecutive amino acid frequencies; 'threeB' = triples + betweenness centrality network properties; 'threeC'=triples + clustering coefficient; 'threeD' = triples + degree (network properties); 'threeS' = triples + sequence similarity feature, and 'threeA' = combination of all the above mention features*

2.4.4 Performance of Quadruple amino acids in combination with network and sequence similarity features

We repeated the procedure of feature combinations, but with triples replaced by quadruples. The results in Table 2.3 show the performance of the quadruple feature and the combinations of quadruple with each of the other features, in order to evaluate the importance of each single feature combined with the quadruples. The results in Table 2.3 represent the model accuracies. The column called *model average* shows the average of

the three numbers in the other columns, i.e., the average of the accuracies of the training, testing and validation. In Table 2.3 we see that the prediction using the quadruple feature alone, gives an accuracy of 70.7%. Combining all the features gives an improvement in accuracy (93.4%). This shows the importance of graph property features and the sequence similarity. Figure 2.3 visualizes the results presented in Table 2.3 by means of ROC curves. Recall that the significance of the ROC curve is such that, as prediction gets closer to 100% accuracy, the ROC curve will tend towards the upper left corner. In this case Figure 2.3 shows that the combination of quadruples with all other features is the most accurate among the predictors. Finally, in the overall comparison of model performance we find that the quadruple feature combined with other features constitutes the best model among those presented in this Chapter. We use this model to make predictions of new human-*Bacillus* PPIs.

Table 2.3: Model performance of quadruple amino acids with combination of other features

Features	Model average	Training	Resting	validation
$four$[1]	70.7	73.5	64.2	64.2
$fourD$[2]	80.1	82.9	76.5	70.6
$fourB$[3]	88.0	90.8	83.3	79.9
$fourC$[4]	91.0	95.4	80.4	80.9
$fourS$[5]	86.4	89.0	81.4	79.4
$fourA$[6]	93.4	97.3	82.8	85.8

[1]*quadruple* consecutive amino acid frequencies.

[2]*quadruples+* betweenness centrality network properties.

[3]*quadruples+* clustering coefficient.

[4]*quadruples+* degree (network properties).

[5]*quadruples+* sequence similarity feature.

[6]*combination* of all the above mentioned features.

ROC curve for quadruple amino acids in combination with other features

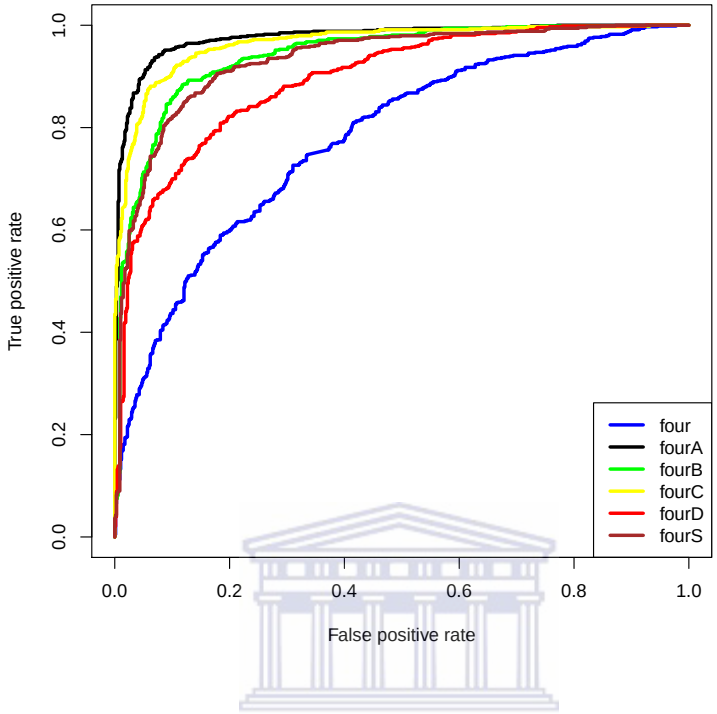


Figure 2.3: ROC curves for six different combinations of feature sets

Each curve indicates the feature set used. 'four' = quadruple of four consecutive amino acid frequencies; 'fourB' = quadruples + betweenness centrality network properties,; 'fourC' = quadruples + clustering coefficient, 'fourD' = quadruples + degree (network properties); 'fourS' = quadruples + sequence similarity feature; 'fourA' = combinations of all above mentioned features.

2.4.5 Prediction of human-*Bacillus* PPIs

The model based on the quadruples feature combined with the sequence similarity, and human interactome graph properties were chosen as an optimal model. We use this model to make predictions of new human-*Bacillus* PPIs (Figure 2.4, 2.6 and 2.6).

2.4.6 Functional enrichment analysis of sub-network

Functional enrichment analysis uses statistical methods to find functions that are over-represented in a subset of genes. Thus it is very important for identifying the functional relevance of proteins involved in the host-pathogen PPIs. The presence of over- represented functional categories that are closely related to immune response, can serve as further support for the validation. The lists of significantly enriched GO terms for molecular function are given in appendix A (Tables A.1 and A.2) respectively for Figures 2.4, 2.5 and 2.6.

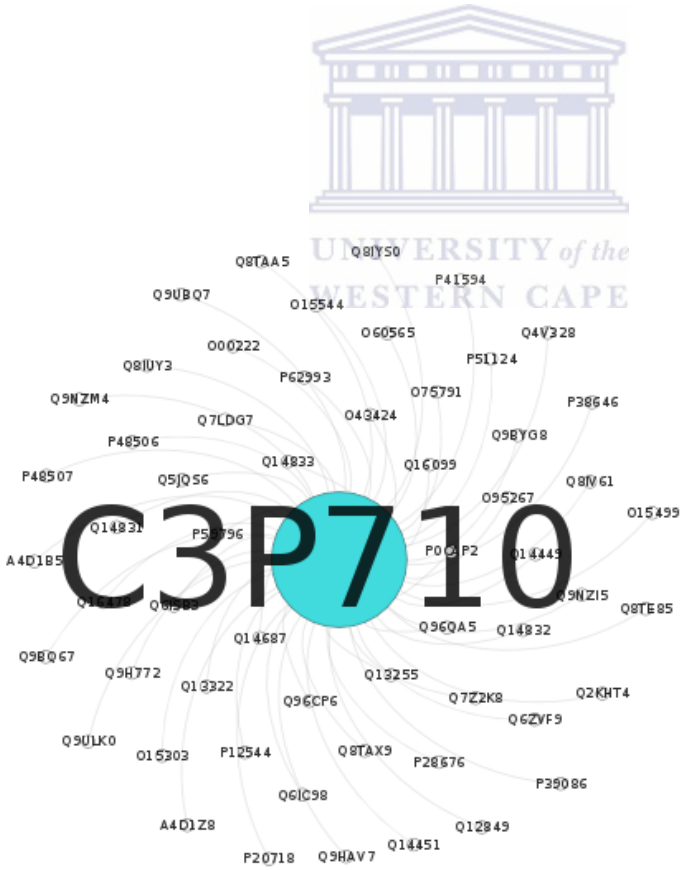


Figure 2.4: Protein interactions predicted between *Bacillus* protein C3P710 and human proteins

Figure 2.5: Protein interactions predicted between *Bacillus* protein C3P8D5 and human proteins

Figure 2.6: Protein interactions predicted between *Bacillus* protein C3P5Q9 and human proteins

2.5 Conclusion

Knowledge of interactions between host and pathogen proteins is important for under standing the pathogenic process. The goal of this study was to define an optimal feature set for, and subsequently, to predict physical protein interactions of *Bacillus anthracis* with human proteins, using a neural network trained with human-*Bacillus anthracis* PPIs data. Different combinations of features were used to test the model performance. The best performance was the model trained with amino acid quadruples and pairwise sequence similarity, together with human interactome properties such as degree, cluster coefficient and betweenness centrality. This confirm that assumption that state pathogens are tend to target hub proteins. Our approach demonstrated that the feature selection was not biased to virus nucleotide composition but could be used in the context of bacterial genomes.

Chapter 3

Human-*Mycobacterium tuberculosis* PPI Prediction

3.1 Abstract

Background: Tuberculosis is one of the most significant infectious diseases affecting humans, caused by *M-tuberculosis*. The lack of effective vaccine and antibiotics, and TB-HIV co-infection with the emergence of both multi drug resistance (MDR) and extreme drug resistance (XDR) make TB a serious global health threat. Studying human-*M-tuberculosis* protein-protein interactions will help to understand the virulence and mechanisms of this pathogen, and will be helpful in identifying potential drug targets. A previous study used the interolog method combined with domain-domain interaction to predict PPI between human and *M-tuberculosis* proteins, and utilized functional annotation for further validation. Another study by Rapanoel et al. (2013) developed an interolog method to predict PPI between human and *M-tuberculosis*, then filtered the result using differentially expressed gene during infection together with known human-*M-tuberculosis* PPI data. **Results**: In this work we propose a novel strategy to over come the lack of experimental PPI data for human-*M-tuberculosis* PPI prediction. Thus, we used a multi-layer feed-forward neural network to predict human-*M-tuberculosis* PPI. In the absence of a large

training data set for human-*M-tuberculosis* interactions, the network was trained using experimentally validated data on human-*Bacillus anthracis* PPIs from the IntAct database, and combinations of six features namely amino acid triplets or quadruplets, pairwise sequence similarity, and human interactome properties including graph-theoretic properties such as degree, cluster coefficient and betweenness centrality. For further validation, a total of 83 human-*M-tuberculosis* PPIs were identified through orthology mapping to five human-pathogen datasets. Our predictor shows an average accuracy of 93.4% when using quadruple and the human interactome features, compared to an average accuracy of 91.3% for amino acid triplets. An examination of the predicted human-*M-tuberculosis* protein interaction network highlighted the enrichment of human immune related genes interacting with *Mycobacterium tuberculosis* cell membrane proteins(p=0.01). Published secretory proteins for *M-tuberculosis* during infection provided another dataset to cross reference our predicted PPI.

3.2 Overview

Tuberculosis (TB) is an infectious disease usually caused by the bacterium *M-tuberculosis*. The bacteria are easily spread through the air from human to human. TB is responsible for 9.4 million new infections annually, and 1.7 million deaths p.a. according to the World Health Organization records (WHO, 2010). One-third of the world population is currently infected with TB, but only 10 percent of people who are infected will become infectious at some time during their lives.

In 2008, globally there were an estimated 440 000 cases of multi drug resistant tuberculosis (MDR-TB) (Huynen et al., 2000). Therefore, a comprehensive view of the causative organism *Mycobacterium tuberculosis* and its interaction with the host promises to provide higher level of insights into the biology of the organism as well as to provide rational strategies for designing therapeutic agents. In particular, studying human-*M-tuberculosis* protein-protein interaction will help to understand the virulence and mechanisms of this pathogen, and identification of potential drug targets. Protein-protein

interactions (PPIs) are key players in biological functioning on the molecular level. A range of online intra-species protein-protein interaction resources are available that include both experimental and/or computational evidence (a comprehensive list can be found at $https://www.hsls.pitt.edu/obrc/index.php?page=protein_protein_interactions$. Inter-species protein- protein interaction predictions are dominated by species for which there is an abundance of experimental data that can be used for training and testing see review by (Zhou and Wong 2013). Machine learning methods were applied to human-viral interaction data (Dyer et al., 2007, 2008, 2011; Cui et al., 2012, Emamjomeh et al., 2014; Barmen et al., 2014) and to a lesser extent to human-bacterial interactions (Dyer et al., 2010, Mazandu and Mulder 2011, Rapanoel et al., 2013). Despite the limited number of experimentally verified human-*M-tuberculosis* interactions, Huo et al (2015) generated a framework for human-*M-tuberculosis* interactions using a interolog sequence similarity-based approach and included no more than 110 experimentally verified protein-protein interaction pairs. These authors combined domain interaction with functional enrichment for predicting PPIs between human and *M-tuberculosis*. Moreover, a homology-based approach was implemented by Zhou et al. (2014) for human *M-tuberculosis* PPIs prediction. Their predicted host-pathogen list was filtered using cellular compartment distribution analysis, disease gene list enrichment analysis, pathway enrichment analysis and functional category enrichment analysis. Similarly, Rapanoel et al (2013) relied on an interolog protein-protein interaction method that relied on intra-species interactions for human obtained from a database of interacting proteins (DIP) and predicted interactions for *M-tuberculosis* and humans (Mazandu and Mulder 2011; Mazandu and Mulder 2012). Filters such as gene expression data was used to reduce false positives (Rapanoel et al., 2013) Despite the absence of large experimental datasets, we revisited a classification method to predict human-*M-tuberculosis* protein-protein interactions based on the improved performance achieved with a neural network approach combined with a selection of features that included quadruplet amino acids (see Chapter 2). In this chapter we tested the utility of using human-human protein interactions and pathogen- pathogen interactions as part of the training data to train the MFNN. Due to the limited number

of experimentally verified *M-tuberculosis-M-tuberculosis* protein-protein interactions, we first tested the approach using intra-species data for *Bacillus anthracis* and human-human PPI. We refer to this step as the proof of concept model. Based on the accuracy measurements obtained for the MFNN we tested the approach using the limited experimentally validated *M-tuberculosis-M-tuberculosis* PPI and human- human PPI datasets. Finally, we validated our resulting PPI predictions using a set of 44 secretory *M-tuberculosis* proteins that were experimentally shown to be released into macrophages.

3.3 Implementation

The architecture involved for the construction of human-*M-tuberculosis* PPI is shown in Figure 3.1. The first step was extract *Bacillus* intra-species PPIs and human intra species data for training the algorithm. In addition we used host-pathogen data namely human-*Bacillus* PPI data which are experimentally identified, for further validation. This is based on the assumption that a model developed using intra-species data can be extended to host- pathogen PPI prediction considering the extent of sequence similarity between the target and template datasets in prediction. After the evaluation of the method on experimentally verified human-*Bacillus* data, we proceeded with a repetition of the same strategy with human and *M-tuberculosis* intra species data for model building (training and testing) and then make use of it for human-*M-tuberculosis* PPI prediction.

3.3.1 Data

All data used in the construction of the positive data set data set was downloaded in 2012 and updated in 2015 from the database IntAct (Henning et al., 2004) and PATRIC (Ioannis et al., 2002).

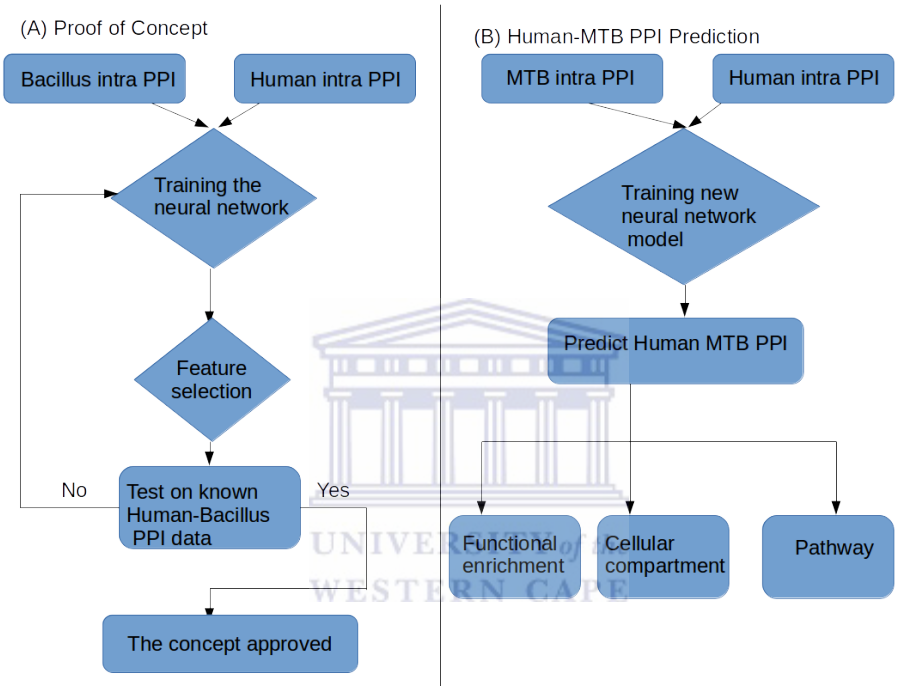


Figure 3.1: Work flow applied to the construction of the human-*Mycobacterium tuberculosis* PPIs predictions. The panel (A) represent the process of the proof of concept model development using human and *Bacillus* intra species PPIs data for model construction, and human-*Bacillus* inter species PPIs data as independent test set; panel (B) Implement the concept for human-*M-tuberculosis* PPIs prediction

3.3.1.1 Proof of Concept Model

For the test model, we collected experimental protein-protein interaction data from IntAct database (Henning et al., 2004) for human to serve as a positive set, which include 200 pairs of interacting proteins. For the second organism *Bacillus anthracis*, there is not enough intra-species experimentally validated PPI data. Therefore, we extracted PPIs for different *Bacillus* from the PATRIC database (Wattam et al., 2014). These species include *Bacillus-amyloliquefaciens-CC178, Bacillus-anthracis-str- A0174, Bacillus-sp-10403023, Bacillus-subtilis-S1-4,*and *Bacillus-subtilis-subsp-natto- BEST195.* A total of 150 pairs of PPIs were collected of Bacilli. All the interacting protein pairs were identified by their UniProtKB (Magrane and UniProt Consortium, 2011) Accession IDs for normalization purposes. In some instances it was necessary to convert the database identifiers to UniProtKB Accession IDs. Thus, the 200 human PPIs and 150 *Bacillus* PPIs were used as positive data sets for training. The selection of a negative data set or non-interacting proteins was identified based on different cellular localization (Ben-Hur and Noble, 2006). These methods consist of randomly selecting protein pairs that are not present in a veto list containing all PPIs from the positive data set. With this strategy we generated a negative set of a size similar to that of the positive sets (200 for human and 150 for *Bacillus* negative protein pairs), and combined it with the positive set to obtain a training data set with 700 PPI pairs. In addition to the training data sets, we downloaded human- *Bacillus anthracis* PPIs data from PATRIC database (Ioannis et al., 2002) for further validation and model generalization.

3.3.1.2 Human *M-tuberculosis* Prediction Model

In order to apply proof of concept approach to human-*M-tuberculosis* PPI predictions, we face the challenge that there is no enough *M-tuberculosis* intra-species PPIs. Consequently we proceed with extracting *Mycobacterium* genus intra-species PPIs from PATRIC database (Wattam et al., 2014). The *Mycobacterium* genus include (*Mycobacterium tuberculosis C, Mycobacterium leprae, Mycobacterium avium 104, Mycobacterium smegmatis*

str. MC2 155, Mycobacterium tuberculosis, Mycobacterium tuberculosis H37Rv). A total of 117 experimentally validated intra-species PPIs for *Mycobacterium* were extracted. This data serve as the positive dataset. On the other hand the interacting pairs for human intra-species were chosen similarly as in Section 2.1.1. There is not enough human-*Mycobacterium tuberculosis* experimentally validated PPI data, therefore we adopt the proof of concept approach for human-*M-tuberculosis* PPI prediction. In the same way as in the proof of concept model the negative set were generated and combined with the positive set to obtain a training data set with 434 PPIs pairs.

3.3.2 Features Selection

In Chapter 2, we defined the optimal feature set to predict PPIs. These include quadruple, sequence similarity and human interactome network properties such as (degree, cluster coefficient and betweenness centrality). The details of feature encoding are provided in Chapter 2 Section 2.2.4.

3.3.3 Feed Forward Neural Network

In this study we use the multi-layer feed-forward neural network (MFFN). The MFFN is popularly used for a wide variety of classification and prediction tasks.

The details of this classifier has been discussed in Chapter 2, see Section 2.3.2.

3.3.4 Performances Evaluation

The human-*Bacillus anthracis* (proof of concept model) classifier performance was evaluated using 3-fold cross validation, on which one third of the examples are reserved for testing. The training data was also further split into three, and 1/3 was used as the validation data. Therefore, we evaluated the quality of our predictive model using the receiver operating characteristic (ROC) curve and confusion matrix, a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied. The curve was created by plotting the true positive rate against the false positive

rate at various threshold settings. It shows the trade off between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity). The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test. The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.

3.3.5 Human-*M-tuberculosis* PPI Validation

Four parameters were used to validate the human-*M-tuberculosis* PPIs in the absence of experimental data namely

(i) functional enrichment analysis.

(ii) cellular compartment distribution.

(v) pathway enrichment analysis.

3.3.5.1 Functional Enrichment Analysis

Functional enrichment analysis is important for identifying the functional relevance of host proteins predicted to be involved in host-pathogen PPIs. The presence of enriched (over-represented) functional categories that are closely related to pathogen infection, serves as further support for the validity of the prediction results. Molecular function term enrichment analysis on the human proteins involved in the predicted human- *M-tuberculosis* PPIs was conducted using the DAVID database (Huang et al., 2009). DAVID does not support the functional enrichment analysis of *M-tuberculosis* proteins and therefore we used an in-house tool to identify the over-represented functional terms for the corresponding *M-tuberculosis* proteins. Chande et al (2015) identified 44 secretory proteins within an infected host macrophage that correspond to the mycobacterial virulent strain (H37Rv). Among these 44 proteins were proteins that function in virulence, detoxification and adaptation. Five of the 44 *M-tuberculosis* secretory proteins were extracted from the human-*M-tuberculosis* PPI dataset to validate the predicted interactions.

3.3.5.2 Cellular Compartment Distribution

The cellular component gene ontology term describes locations, at the levels of sub cellular structures and macromolecular complexes. Examples of cellular components include nuclear inner membrane, with the synonym inner envelop, and the ubiquitin ligase complex, with several subtypes of these complexes represented. Generally, a gene product is located in or is a subcomponent of a particular component. However, the cellular compartment of the human proteins targeted by the predicted host-pathogen PPIs are an important indicator of the quality of PPI prediction. If the targeted human proteins are mostly located in cellular components having a close relationship with pathogen infection or known interactions with host cells that are relevant to the pathogen infection, then we can be more certain about the quality of our prediction. Gene Ontology cellular compartment is one of the most inclusive annotations for human proteins. The cellular compartment distribution shows how many proteins (and the percentage) in the dataset happen to fall into each cellular compartment. We selected the top 20 most frequently located cellular compartments of the human proteins that are predicted to be targeted by *M-tuberculosis* in our model.

3.3.5.3 Pathway Enrichment Analysis

Pathway enrichment analysis is a primary source for identifying a list of functionally related proteins. Therefore, for a set of proteins that are significantly enriched in certain pathways, it is very likely that this set of proteins play coordinated roles in vivo. Thus, pathway enrichment analysis is one of the most frequently used assessments for predicted host-pathogen PPIs. For pathway enrichment analysis, we use DAVID database (Huang et al., 2009), which is currently one of the most extensive integrated enrichment analysis databases. For each human protein set predicted by our model, we analyzed the human proteins pathway enrichment using the DAVID database, and the top 20 most significantly enriched pathways. The enrichment analysis results provide important evidence that our approach can predict more human-*M-tuberculosis* PPIs that are more

relevant to *M-tuberculosis* infection. Besides assessing the quality of the host proteins that are predicted to interact with pathogen proteins based on pathway enrichment, we also conduct pathway enrichment analysis for *M-tuberculosis* proteins that target human pro- teins. This analysis was done by using IntPath database (Zhou et al., 2012) which support pathway enrichment analysis of this important pathogen. The pathway analysis on the *M-tuberculosis* proteins are also used to assess the performance accuracy of our model, which give clues to the functional roles of *M-tuberculosis* proteins that target human proteins.

3.4 Results and Discussion

We develop a test model (proof of concept) using experimentally verified intra species PPI for human-*Bacillus* combined with our feature set described in Chapter 2. After validating our approach using human-*Bacillus*, we proceeded to predict the protein-protein interaction (PPIs) between human and *Mycobacterium tuberculosis*.

3.4.1 Construction of the Proof of Concept Model

Figure 3.1 (a) summarizes the procedure used to construct the model of the human-*Bacillus anthracis* PPI prediction. The starting point of this work is a set of 150 pairs of human protein-protein interaction data extracted from IntAct database (Henning et al., 2002) which serve as a positive set. Since there is no well-established gold standard PPI data for *Bacillus anthracis*, we collected data from PATRIC databases (Wattam et al., 2014) containing four different *Bacillus* high-quality experimentally determined interactions PPIs as described further in Section 3.3.1.1. The data extracted were merged, creating our gold standard of positive interactions. The gold standard of negative interactions was obtained by randomly pairing the protein list from non-interacting protein sets. These randomly generated protein pairs were filtered to ensure that there were no protein pairs that are known to interact in the positive dataset. The final training data sets contain 150 pairs of human PPI positive and negative data, and similarly 150 pairs

for *Bacillus*. Moreover, for each possible pair of proteins, we constructed two type of features based on:(I) pairwise sequence similarity, and; (II) quadruple consecutive amino acid. In addition to human PPI network graph properties values such as (I) degree; (II) betweenness centrality; (III) cluster coefficient; as described in Section 3.3.1.1. The gold standard dataset was used to train a feed forward artificial neural network classifier and to perform further validations on the final model. The proof of concept model achieved an average accuracy of 91.4% on training (Figure 3.2), 85.6% validation (Figure 3.4) and 80.6% testing (Figure 3.5). In addition, the overall average accuracy of training, testing and validation was 89.0% (Figure 3.3) which indicate good promise for transferring the model to host-pathogen PPI prediction. Thus, in order to make use of the model developed on intra-species data for host-pathogen interaction, we test our model on independent human-*Bacillus anthracis* PPI data. The independent test sets consist of 680 pairs of human-*Bacillus anthraces* PPIs experimentally identified PPI data extracted from IntAct database (Henning et al, 2002). Finally, the result of the independent test is used to generalize the concept of training a classifier with intra-species PPIs, and apply it to the problem of host-pathogen PPIs prediction.

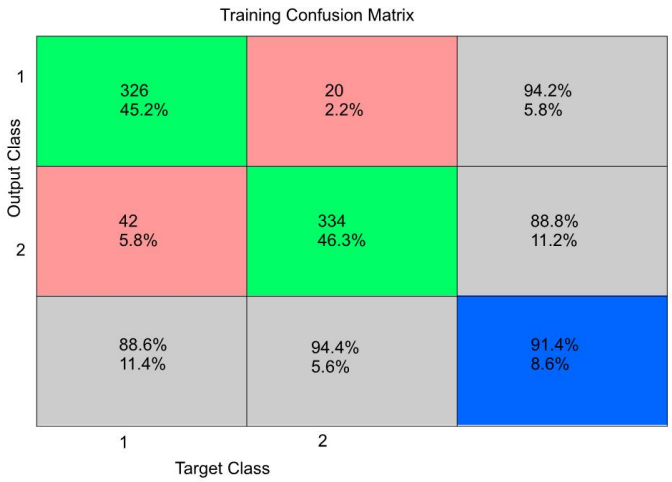


Figure 3.2: Training confusion matrix for the proof of concept model

Training confusion matrix for human-*Bacillus* PPI. In the confusion matrix plot the rows show the predicted class and the columns show the true class. The diagonal cells show where the true class and predicted class match. The off diagonal cells show instances where the classifier has made mistakes. The column on the right hand side of the plot shows the accuracy for each predicted class, while the row at the bottom of the plot shows the accuracy for each true class. The cell in the bottom right of the plot shows the overall accuracy.

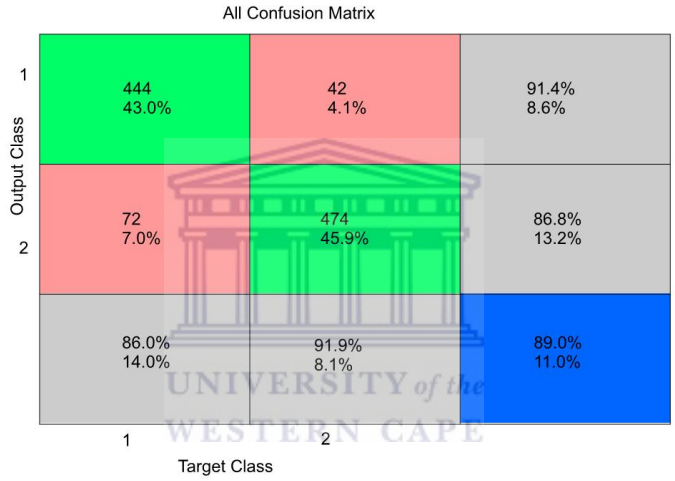


Figure 3.3: Confusion matrix plot that reflect the average result of training, testing and validation process for the proof of concept model

3.3 Confusion matrix plot that reflect the average result of training, testing and validation process for the human-*Bacillus* PPI: In the matrix the rows show the predicted class and the columns show the true class. The diagonal cells show where the true class and predicted class match. The off diagonal cells show instances where the classifier has made mistakes. The column on the right hand side of the plot shows the accuracy for each predicted class, while the row at the bottom of the plot shows the accuracy for each true class. The cell in the bottom right of the plot shows the overall accuracy (blue).

3.4.2 human-Bacillus anthracis PPI Performance Evaluation

In this section, we evaluate the performance of the proposed method when applied to the set of human and *Bacillus* intra-species PPI data respectively. We performed a 3-fold cross-validation to assess the model performance. Cross-validation is a model validation technique for assessing how the results of a statistical analysis will generalize to an independent data set. Figure 3.2, 3.4 and 3.5 shows the confusion matrix of the training, validation and testing respectively. The confusion matrix is represented by a matrix each row of which represents the instances in a predicted class, while each column represents the actual class. One of the advantages of using this performance evaluation tool is that the data mining analyzer can easily see whether or not the model is confusing two classes. The matrix also shows the accuracy of the classifier as the percentage of correctly classified patterns in a given class divided by the total number of patterns in that class. The overall (average) accuracy of the classifier is also evaluated by using the confusion matrix Figure 3.3. In addition to the confusion matrix we use the ROC curve to illustrate the performance of the classifier. The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. Figure 3.6 represent the training result for the proof of concept model. Figure 3.7 shows the proof of concept model performance result on independent host-pathogen data for human-*Bacillus anthraces* PPIs data that are experimentally verified.

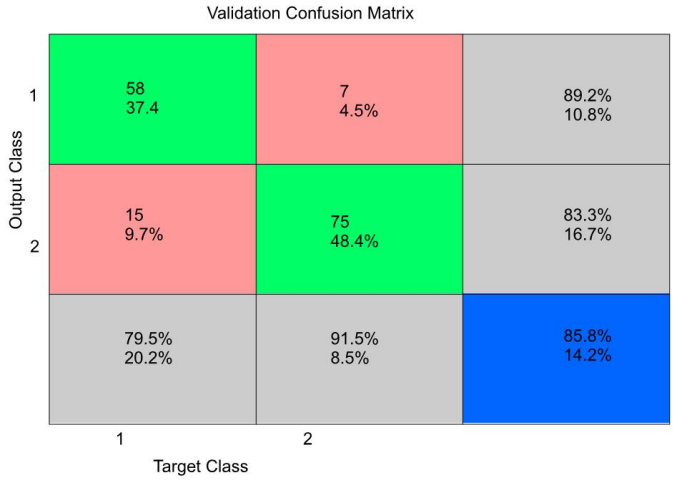


Figure 3.4: Validation result for the proof of concept model

In the matrix the rows show the predicted class and the columns show the true class. The diagonal cells show where the true class and predicted class match. The off diagonal cells show instances where the classifier has made mistakes. The column on the right hand side of the plot shows the accuracy for each predicted class, while the row at the bottom of the plot shows the accuracy for each true class. The cell in the bottom right of the plot shows the overall accuracy.

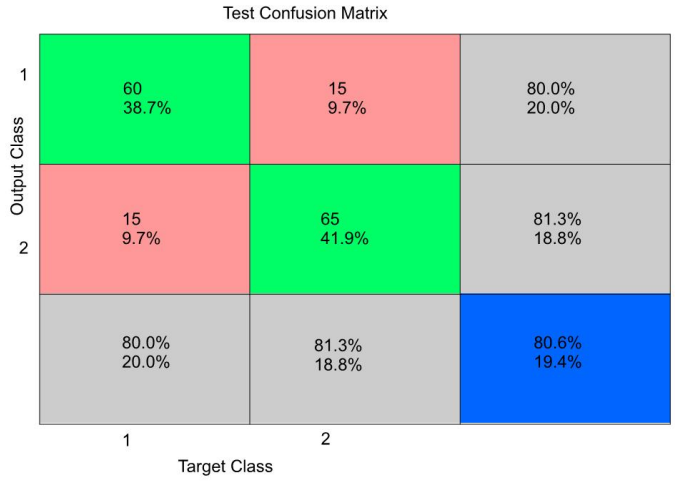


Figure 3.5: Proof of concept model testing result

3.5 Validation result for the human-*Bacillus* PPI: In the matrix the rows show the show the predicted class and the columns show the true class. The diagonal cells show where the true class and predicted class match. The off diagonal cells show instances where the classifier has made mistakes. The column on the right hand side of the plot shows the accuracy for each predicted class, while the row at the bottom of the plot shows the accuracy for each true class. The cell in the bottom right of the plot shows the overall accuracy (blue).

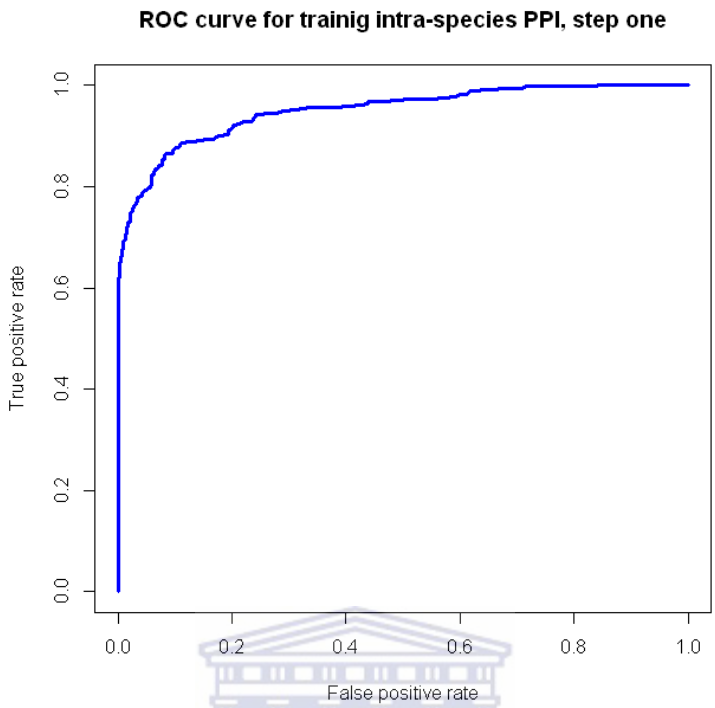


Figure 3.6: ROC curve of proof of concept model (proof of concept model)

The curve represent the accuracy of the model using six different combinations of feature sets, namely quadruples consecutive amino acid frequencies, betweenness centrality, clustering coefficient, degree, and sequence similarity.

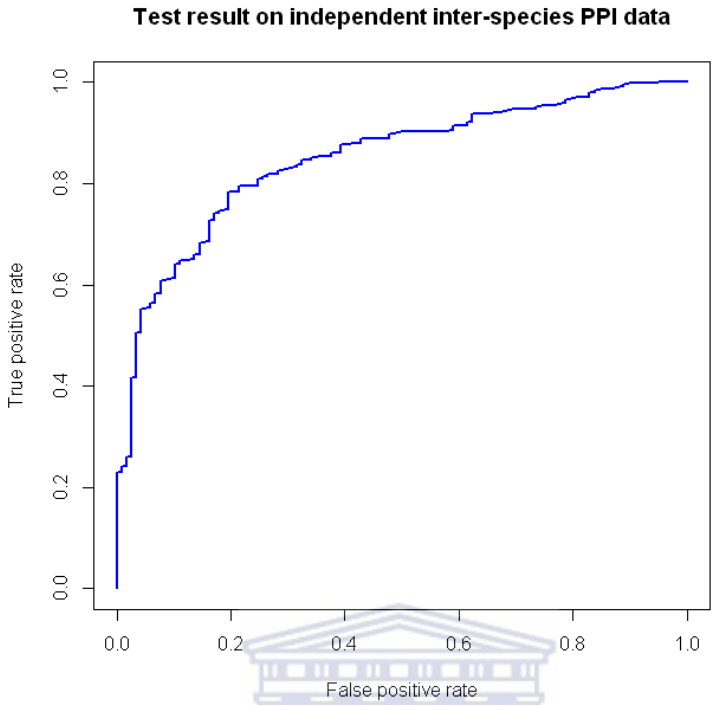


Figure 3.7: ROC curve of proof of concept model (proof of concept model)

The curve represent the accuracy of the model on independent human-*Bacillus anthraces* PPIs data that has been experimentally verified.

3.4.3 Construction of Human-*Mycobacterium tuberculosis* PPI Prediction Model

In the previous section we demonstrated a model that tests the hypotheses which starts by considering intra-species protein-protein interactions (especially intra human and intra *M-tuberculosis*) as the training set, and then we consider the extent of sequence similarity between the target and template datasets in prediction. We follow the same procedure as mentioned in the test model construction section 3.4.1.

The same problem formulation as described in the previous section for the earlier model, was applied to develop the new model described in this section. Predicting physical interactions between human and *Mycobacterium tuberculosis* protein pairs is considered as

a binary classification task. That is, each human-*M-tuberculosis* protein pair belongs to one of two classes: interaction or non-interaction. Associated with every protein, is a numeric feature vector. Using labeled examples of the two classes and the feature vectors, a function that distinguishes the two classes is learned using the neural network classifier (see Chapter 2 Section 2.3.2 for details). The human positive and negative sets employed were the same as in the previous model presented in the testing model, see section 3.3.1. In the case of *Mycobacterium tuberculosis* there is not enough gold standard positive interactions data available. However, we extracted protein-protein interaction data from PATRIC database (Wattam et al, 2014). The data sets contain different strains of *Mycobacterium* such as (*Mycobacterium tuberculosis C, Mycobacterium leprae, Mycobacterium tuberculosis, Mycobacterium tuberculosis H37Rv*). After filtering the data obtained from these strains, we generate 150 pairs of interacting *M-tuberculosis* proteins as a positive set. In addition, the negative dataset was generated randomly as described in Section 3.31.2. The total positive set included 500 proteins involving 150 *Mycobacterium tuberculosis* proteins and 200 human proteins. In the absence of a comprehensive negative training set of non-interacting protein pairs, a simple heuristic was applied. As a negative, non-interacting training set of equal size, protein pairs that did not appear in the positive training set were randomly sampled. Using such sets, the feed forward neural network algorithm was applied, allowing cross-validation by reporting the fraction of protein pairs that were correctly classified. The training process was repeated choosing different network structure until the best performance was achieved, returning an area under the ROC curve, (Figure 3.8). This result is above the performance of the test model constructed in the previous section 3.3.1. In addition to the ROC, the confusion matrix was also used for further assessment of the training process. Figures 3.9, 3.10, 3.11, and 3.12 show the training, validation, testing, and the overall average model performance. For example, in the training stage the model achieved 96.6% for training, 84.5% validation, 91.4% testing and 94.0% overall model average.

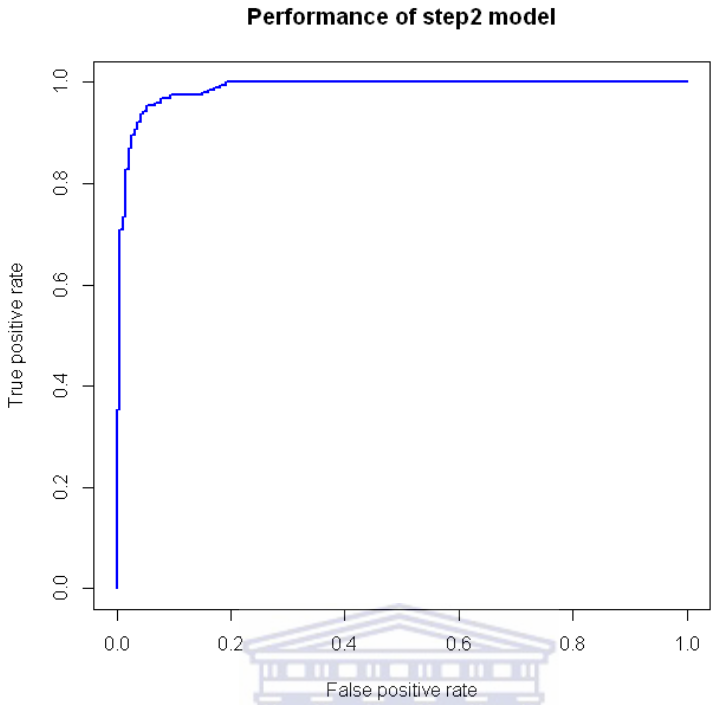


Figure 3.8: ROC curve for Human-*Mycobacterium tuberculosis* model

The curve represents the accuracy of the model using six different combinations of feature sets.

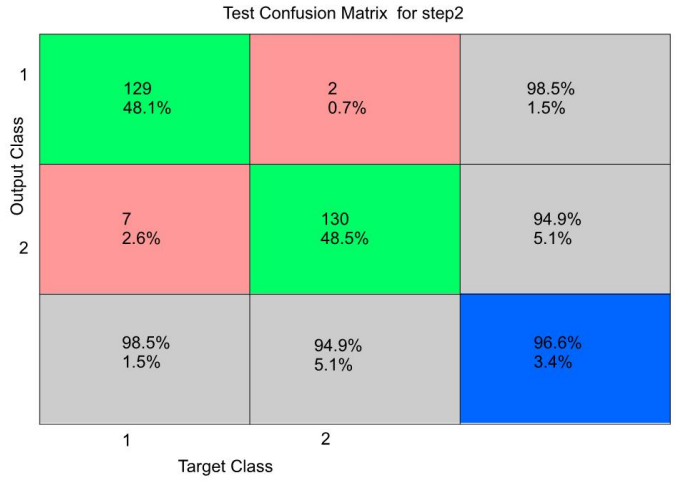


Figure 3.9: Training confusion matrix for the Human-*Mycobacterium tuberculosis* model

3.9 **Training confusion matrix for the Human-*Mycobacterium tuberculosis* model (Step2 model)** In the confusion matrix plot the rows show the predicted class, and the columns show the true class. The diagonal cells show where the true class and predicted class match. The off diagonal cells show instances where the classifier has made mistakes. The column on the right hand side of the plot shows the accuracy for each predicted class, while the row at the bottom of the plot shows the accuracy for each true class. The cell in the bottom right of the plot shows the overall accuracy.

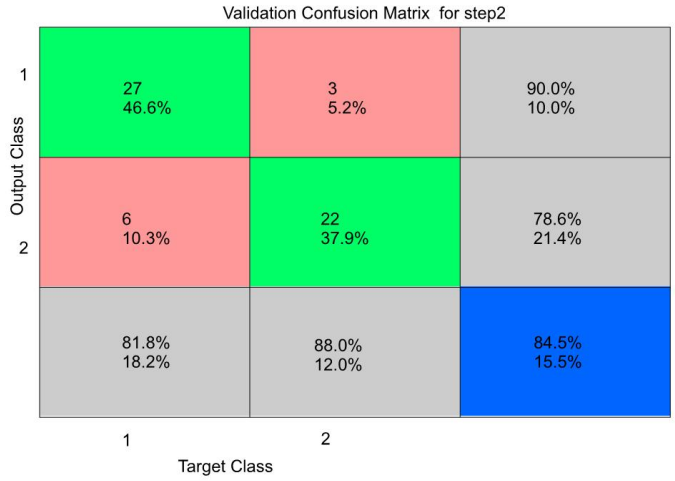


Figure 3.10: Validation result for the Human-*Mycobacterium tuberculosis* model

3.10 **Validation result for the Human-*Mycobacterium tuberculosis* model (Step2 model).** In the matrix the rows show the predicted class, and the columns show the true class. The diagonal cells show where the true class and predicted class match. The off diagonal cells show instances where the classifier has made mistakes. The column on the right hand side of the plot shows the accuracy for each predicted class, while the row at the bottom of the plot shows the accuracy for each true class. The cell in the bottom right of the plot shows the overall accuracy.

Therefore, the model described above was used for the prediction host-pathogen protein-protein interaction, namely human as host and *M-tuberculosis* as pathogen. We start off the prediction process by preparing blind sets of unknown human-*M-tuberculosis* interaction sets. The classifier returned a set 7750 human proteins that interact with 1171 *M-tuberculosis* proteins (Appendix Table S3,S4 supplementary data).

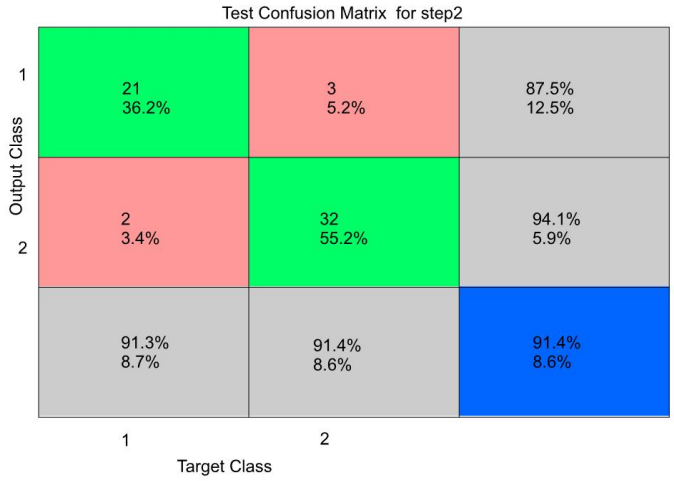


Figure 3.11: The Human-*Mycobacterium tuberculosis* model

3.11 **The testing result for the Human-*Mycobacterium tuberculosis* model (Step2 model)**. In the confusion matrix the rows show the predicted class, and the columns show the true class. The diagonal cells show where the true class and predicted class match. The off diagonal cells show instances where the classifier has made mistakes. The column on the right hand side of the plot shows the accuracy for each predicted class, while the row at the bottom of the plot shows the accuracy for each true class. The cell in the bottom right of the plot shows the overall accuracy.

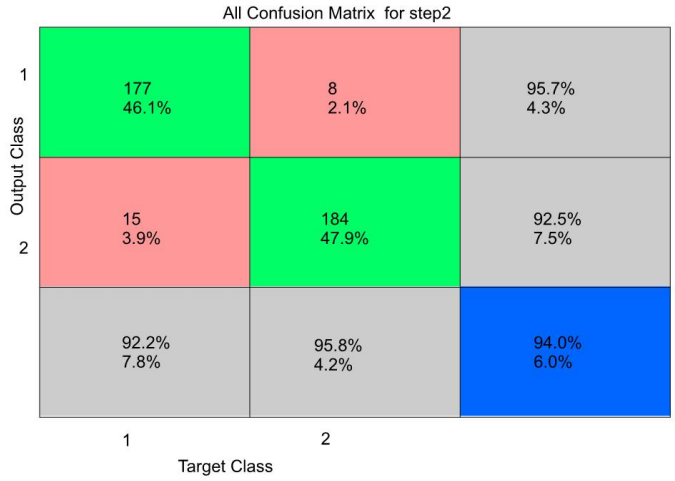


Figure 3.12: The confusion matrix plot that reflect the average result of (training, testing and validation) process for the Human-*Mycobacterium tuberculosis* model

3.12 **The confusion matrix plot that reflect the average result of (training, testing and validation) process for the Human-*Mycobacterium tuberculosis* model (Step2 mode).** In the matrix, the rows show the predicted class and the columns show the true class. The diagonal cells show where the true class and predicted class match. The off diagonal cells show instances where the classifier has made mistakes. The column on the right hand side of the plot shows the accuracy for each predicted class, while the row at the bottom of the plot shows the accuracy for each true class. The cell in the bottom right of the plot shows the overall accuracy.

3.4.4 Quality Assessment of Candidate Human Proteins Predicted to Interact with *M-tuberculosis*

While the neural network classifier provided a large sample of candidates, such interactions most certainly contained a considerable amount of false positives. To assess the quality of interactions we use various items of biological evidence for further filtering. Such evidence include gene ontology terms enrichment analysis, pathway analysis and network

topology. This result shows a promising strategy for overcoming the lack of training data for driving a supervised classifier, where the PPIs data was introduced from different strain of *Mycobacterium* species and incorporating appropriate feature sets. The functional assessment of the PPI network demonstrated that the interacting proteins are involved in immunity-related functions and provide clues to the role of hypothetical proteins in *M-tuberculosis*.

3.4.4.1 Functional Enrichment Analysis

Functional enrichment analysis is important for identifying the functional relevance of host proteins predicted to be involved in host-pathogen PPIs. The presence of enriched (over-represented) functional categories that are closely related to pathogen infection, serves as further support for the validity of the prediction results. Molecular function term enrichment analysis on the human proteins involved in the predicted human- *M-tuberculosis* PPIs was conducted using the DAVID database (Huang et al., 2009) Table A.3 appendix A. DAVID does not support the functional enrichment analysis of *M-tuberculosis* proteins and therefore we used an in-house tool to calculate the over-represented functional terms for the corresponding *M-tuberculosis* proteins.

Chande et al (2015) identified 44 secretory proteins within an infected host macrophage that correspond to the mycobacterial virulent strain (H37Rv). Among these 44 proteins were proteins that function in virulence, detoxification and adaptation. Five of the 44 *M-tuberculosis* secretory proteins were extracted from the human-*M-tuberculosis* PPI dataset to validate the predicted interactions. The chaperon protein, P9WPE9, is predicted to interact with 34 human proteins Figure 3.14. Functional classes identified for the interacting human proteins are oxidoreductase activity, acting on NAD(P)H, oxygen as acceptor, interferon binding, tetrapyrrole binding and heme binding (3.5). This data supports previous work that showed that chaperons facilitate efficient mycobacterial association with macrophages and polarizing in M2-like phenotype (Hickey et al 2014; Lopes et al 2014). The phospodiesterase cdA (P9WP65) is predicted to interact with 110 human proteins Functional classes identified for the interacting human proteins is MHC class II receptor

activity. This data supports 3.1. The three *M-tuberculosis* proteins identified above are depicted in an interaction network with shared proteins (Figure 3.15) Figure 3.13.

Table 3.1: Functional enrichment analysis of human proteins interaction with P9WP65 *M-tuberculosis*

GO Term ID	GO Term Name	Corrected p-val
GO:0032395	MHC class II receptor activity	7.5E-12
GO:0032393	MHC class I receptor activity	1.8E-2

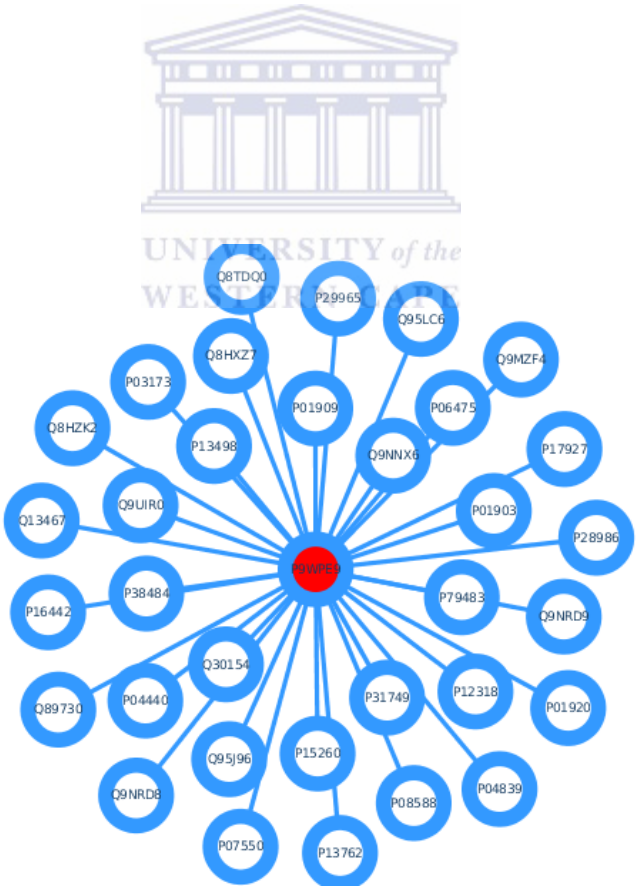


Figure 3.13: A subnetwork of *Mycobacterium tuberculosis* P9WPE9 protein predicted with 34 human proteins.

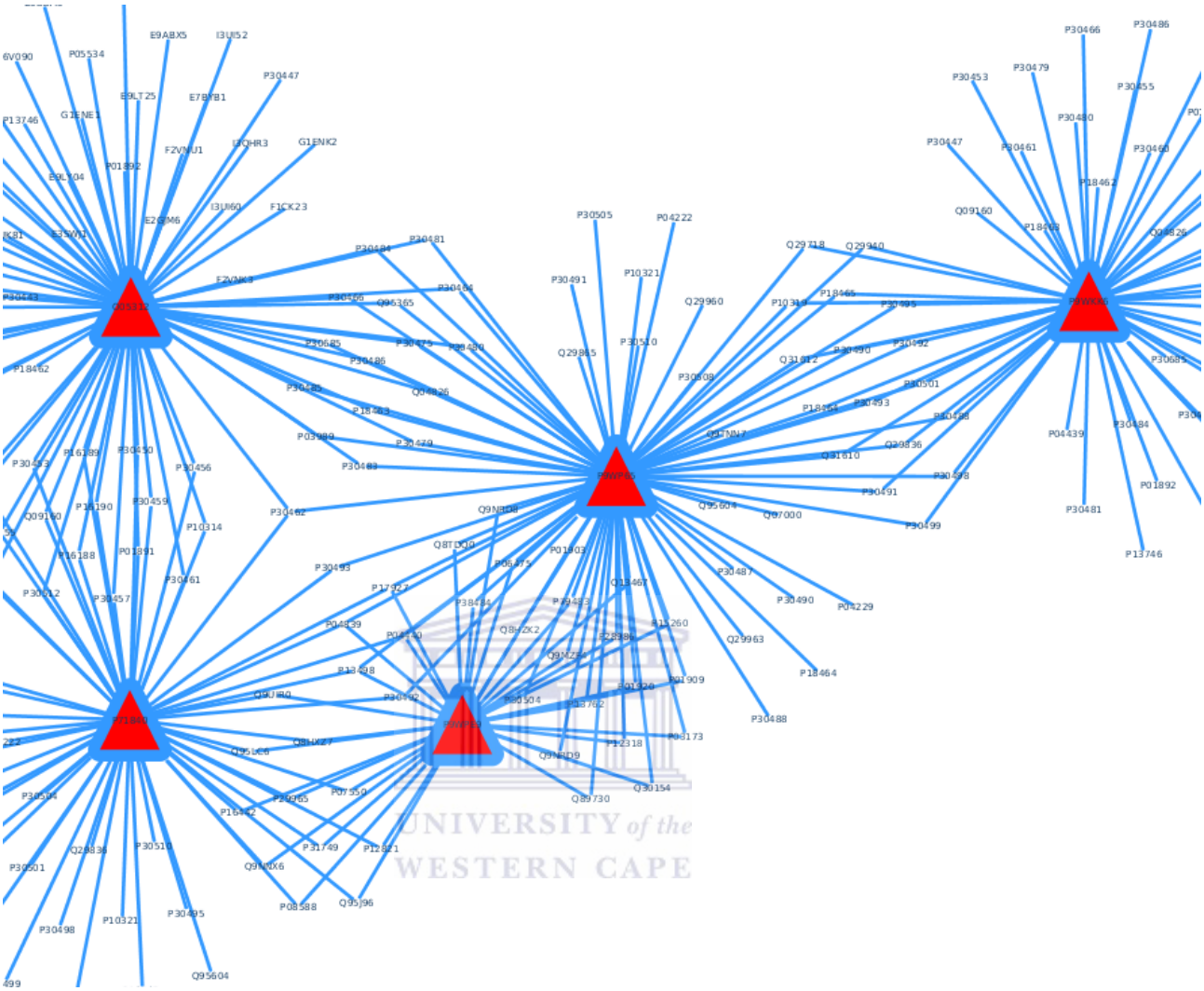


Figure 3.14: A subnetwork of predicted interactions between human-*Mycobacterium tuberculosis* PPI.

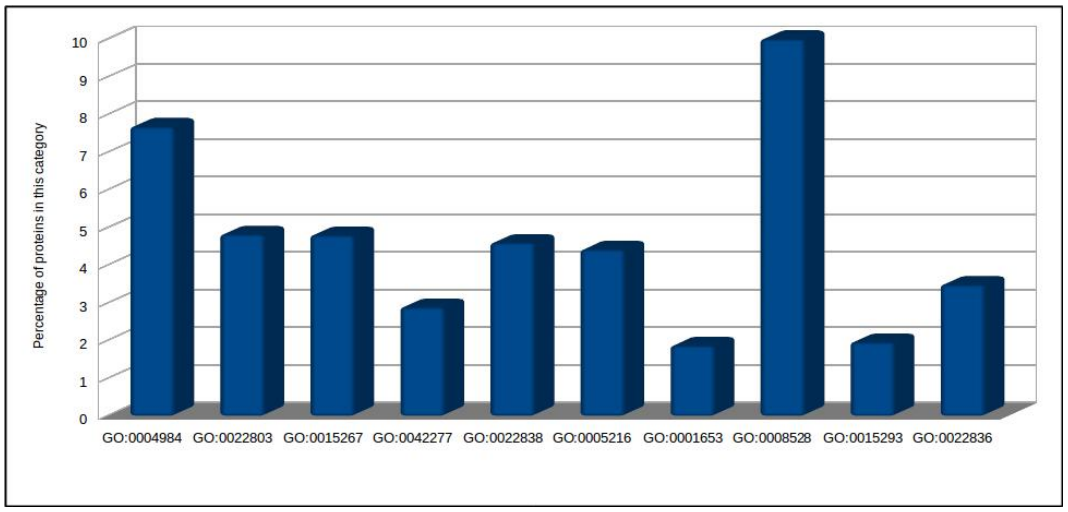


Figure 3.15: Molecular function distribution of human proteins targeted by *M-tuberculosis* predicted by our model.

Table 3.2: Functional enrichment analysis of TB proteins involved in the predicted human-*Mycobacterium tuberculosis* PPIs dataset part1

GO Term ID	GO Term Name	Corrected p-val	Uncorrected p-val
GO:0009408	response to heat	4.50488939444e-09	3.36185775705e-11
GO:0005618	cell wall	1.39097140604e-05	2.07607672543e-07
GO:0006457	protein folding	0.0001357126	3.03834186879e-06
GO:0004451	isocitrate lyase activity	0.0007525261	2.48198542211e-05
GO:0001666	response to hypoxia	0.0007525261	2.80793311456e-05
GO:0009405	pathogenesis	0.00140295	8.28319626094e-05
GO:0005829	cytosol	0.00140295	8.37582093625e-05
GO:0006097	glyoxylate cycle	0.0021493251	0.0001443577
GO:0052572	response to host immune response	0.0039852804	0.000297409
GO:0046677	response to antibiotic	0.0070477864	0.0005785496
GO:0046421	methylisocitrate lyase activity	0.0073154356	0.0007097064
GO:0042026	protein refolding	0.0073154356	0.0007097064
GO:0042542	response to hydrogen peroxide	0.0104922548	0.0011745061
GO:0071451	cellular response to superoxide	0.0104922548	0.0011745061
GO:0009410	response to xenobiotic stimulus	0.0130229335	0.0017493493
GO:0006102	isocitrate metabolic process	0.0130229335	0.0017493493
GO:0046812	host cell surface binding	0.0162933934	0.0024318498

Table 3.3: Functional enrichment analysis of TB proteins involved in the predicted human-*Mycobacterium tuberculosis* PPIs dataset part2

GO Term ID	GO Term Name	Corrected p-val	Uncorrected p-val
GO:0003899	DNA-directed RNA polymerase activity	0.0162933934	0.0024318498
GO:0005886	plasma membrane	0.0196106491	0.0031354111
GO:0071456	cellular response to hypoxia	0.0196106491	0.0032196588
GO:0006099	tricarboxylic acid cycle	0.0199714431	0.0034279343
GO:0040007	growth	0.0226579541	0.004058141
GO:0010039	response to iron ion	0.0262948777	0.0051019912
GO:0009267	cellular response to starvation	0.0262948777	0.0051019912
GO:0051701	interaction with host	0.0296699791	0.0059782794
GO:0001101	response to acid chemical	0.0400090554	0.0086586762
GO:0006352	DNA-templated transcription, initiation	0.0438236319	0.0100310406
GO:0005737	cytoplasm	0.0438236319	0.0101383029
GO:0016987	sigma factor activity	0.048128093	0.0114932759
GO:0005576	extracellular region	0.0495032259	0.012191093
GO:0004601	peroxidase activity	0.0536758885	0.0163986954
GO:0051409	response to nitrosative stress	0.0536758885	0.0182000653
GO:0033670	regulation of NAD+ kinase activity	0.0536758885	0.0220311483
GO:0006534	cysteine metabolic process	0.0536758885	0.0220311483
GO:0015038	glutathione disulfide oxidoreductase activity	0.0536758885	0.0220311483
GO:0050440	2-methylcitrate synthase activity	0.0536758885	0.0220311483
GO:0051336	regulation of hydrolase activity	0.0536758885	0.0220311483
GO:0004096	catalase activity	0.0536758885	0.0220311483
GO:0036440	citrate synthase activity	0.05367588857	0.0220311483
GO:0042744	hydrogen peroxide catabolic process	0.0536758885	0.0220311483
GO:0000774	adenyl-nucleotide exchange factor activity	0.0536758885	0.0220311483
GO:0010034	response to acetate	0.0536758885	0.0220311483
GO:0006880	intracellular sequestering of iron ion	0.0536758885	0.0220311483
GO:0070301	cellular response to hydrogen peroxide	0.0536758885	0.0220311483

Table 3.4: Functional enrichment analysis of TB proteins involved in the predicted human-*Mycobacterium tuberculosis* PPIs dataset part3

GO Term ID	GO Term Name	Corrected p-val	Uncorrected p-val
GO:0080007	S-nitrosoglutathione reductase activity	0.0536758885	0.0220311483
GO:0047547	2-methylcitrate dehydratase activity	0.0536758885	0.0220311483
GO:0031071	cysteine desulfurase activity	0.0630585067	0.0328680353
GO:0003994	aconitate hydratase activity	0.0630585067	0.0328680353
GO:0006826	iron ion transport	0.0630585067	0.0328680353
GO:0090143	nucleoid organization	0.0630585067	0.0328680353
GO:0008260	3-oxoacid CoA-transferase activity	0.0630585067	0.0328680353
GO:0044183	protein binding involved in protein folding	0.0630585067	0.0328680353
GO:0034605	cellular response to heat	0.0630585067	0.0328680353
GO:0043175	RNA polymerase core enzyme binding	0.0630585067	0.0328680353
GO:0015771	trehalose transport	0.0630585067	0.0328680353
GO:0044406	adhesion of symbiont to host	0.0630585067	0.0328680353
GO:0006414	translational elongation	0.0630585067	0.0328680353
GO:0070542	response to fatty acid	0.0630585067	0.0328680353
GO:0046777	protein autophosphorylation	0.0630585067	0.032941011
GO:0006021	inositol biosynthetic process	0.0758535336	0.0435874783
GO:0070404	NADH binding	0.0758535336	0.0435874783
GO:0015968	stringent response	0.0758535336	0.0435874783
GO:0097691	bacterial extracellular vesicle	0.0758535336	0.0435874783
GO:0008199	ferric iron binding	0.0758535336	0.0435874783
GO:0004322	ferroxidase activity	0.0758535336	0.0435874783
GO:0042262	DNA protection	0.0758535336	0.0435874783
GO:0098869	cellular oxidant detoxification	0.0896488476	0.0541907213
GO:0015036	disulfide oxidoreductase activity	0.0896488476	0.0541907213
GO:0050708	regulation of protein secretion	0.0896488476	0.0541907213
GO:0016485	protein processing	0.0896488476	0.0541907213

3.4.4.1.1 Pathway Enrichment Analysis of Proteins Involved in Host-Pathogen PPIs Pathway enrichment analysis of human protein predicted to be targeted by *M-tuberculosis* can reveal much about the functional relevance of host proteins involved in the host-pathogen PPI. The basis for the pathway enrichment analysis stems from the fact 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 infection. We conducted pathway enrichment analysis to assess the quality of our prediction results (Table 3.5) and (Table A.5 in appendix) .

Table 3.5: Pathway enrichment analysis of human proteins involved in the predicted host-pathogen PPIs dataset

GO Term	Description	corrected p-Value
hsa05330	Allograft rejection	0.0655169228523671
hsa04940	Type I diabetes mellitus	0.0587534383209521
hsa05332	Graft-versus-host disease	0.0573589240685808
hsa04620	Toll-like receptor signaling pathway	0.0522176682089858
hsa05320	Autoimmune thyroid disease	0.0411967332397079
hsa04060	Cytokine-cytokine receptor interaction	0.0127952357982488
hsa05310	Asthma	0.0116853080549702
hsa04672	Intestinal immune network for IgA production	0.00852708916093116
hsa04612	AnSystemic lupus erythematosus	0.00600444889205232

3.5 Summary

Knowledge of interactions between host and pathogen proteins is important for understanding the pathogenic process. The goal of this study was prediction of physical interactions of proteins of *Mycobacterium tuberculosis* with human proteins, using a trained neural network. We proposed a novel strategy for host-pathogen PPI prediction in the absence of experimentally validated data. This strategy utilized tow intra-species PPI

data to develop a binary classification model then extend the model for host-pathogen prediction. Therefore, we start with proof of concept model that is built using human intra PPI data combined with *Bacillus anthracis* intra PPI data. Thus, the model was trained using human and *Bacillus* intra species data combined with the optimal feature sets that obtained from Chapter 2. The model was tested using experimentally verified human-*Bacillus anthracis* inter-species PPI data to validated the possibility of extending the model developed using intra species data to make prediction on inter-species PPI. The rationale behind using human-*Bacillus* to build the proof of concept model is that, first there is enough human and *Bacillus* inter species PPI data which is important for testing the model. Secondly, to our knowledge there is no previous work that implemented this strategy for host-pathogen PPI prediction using machine learning techniques so that we can conduct any comparison. However, our model shows good results for inter-species data that motivate us to proceed for developing a model to predict PPI between human and *Mycobacterium tuberculosis*. The human-*M-tuberculosis* PPI prediction results were further filtered using functional annotation categories.

Chapter 4

Online Human-*M-tuberculosis* PPI Predictor

4.1 Abstract

Background: A number of web tools have been developed to predict human-pathogen protein-protein interactions that are based on homology searching methods. These include structural information obtained from interacting domains. Despite the use of machine-learning methods to prediction PPI, there has not been a web-accessible platform that allows a user to screen their data against these classifier models. **Results**: In this study, a host-pathogen predictor web server was designed to allow the user to submit protein sequence pairs for human and *M-tuberculosis*. The front end of the server was written in PHPframework, CSS and javascript, and the back-end program for protein-protein interaction prediction. The PPI prediction module comprises a protein sequence data pre-processing step and a machine learning algorithm for binary classification. The classification algorithm was implemented using pybrain, a python library for artificial neural networks. The HPPredict webserver calculates the likelihood of a human-*M-tuberculosis* protein-protein interaction using an underlying neural network model, which performs with an accuracy of 93% as demonstrated in Chapter 3. The server can be accessed at

URL:(http://hppredict.sanbi.ac.za).

4.2 Overview

To complement experimental data, several computational tools have been developed to predict PPIs within single species (Intra) and between host-pathogen (Inter species). However, only few online tools are available for PPI prediction. Chinnasamy et al. (2006) developed a probability-based tree augmented naive (TAN) Bayesian network combined with yeast PPI data for training and validation to predict protein-protein interactions within a species. In addition, Aloy and Russell, (2003) produced a web server to predict PPIs using homologous searching methods. Their method starts with searching homologs of query proteins against the database of interacting domains (DBID) of known three-dimensional complex structures. The preservation of the atomic contacts at the interaction interface was used as a scoring matrix. Structure information has been central to number of PPI prediction webtools (Ogmen et al., 2005; Planas-Iglesias et al., 2013). Dohkan et al. (2006) developed a web server using a support vector machine (SVM) model to predict PPI in yeast and human. They assume that the high level of false positives in binary classification is due to the equality between positive and negative training sets. Therefore, in order to improve the performance they increased the number of negatives. Rashid et al. (2010) developed a web-server to predict PPIs in *Mycobacterium tuberculosis*. They used an SVM combined with three models: i) amino acid composition, ii) dipeptide composition and iii) biochemical class tripeptide composition. Herein, we developed a web-based tool (HPPredict) to predict potential PPIs between human and *Mycobacterium tuberculosis* that uses a feed forward artificial neural network method to decide whether two proteins interact or not. Quadruple frequency of amino acid, sequence similarity, and human interactome network properties were added as features. Furthermore, HPPredict provides a likelihood score for the potential predicted PPIs.

4.3 Implementation

HPPredict consists of two parts; a front-end web interface, written in PHPframework, CSS and javascript, and a back-end program for protein-protein interaction prediction which consists of two modules. One module focuses on protein sequence data pre-processing, and a machine learning algorithm for binary classification. The classification algorithm was implemented using pybrain, a python library for artificial neural network. The flowchart representation of HPPredict is shown in 4.1. The back-end processes start by data parsing where the input sequences are saved in a temporary file with job ID as filename which is unique for each job having been submitted to the server. Secondly, the data encoding processes are started by calling a number of python scripts This includes a script for the five types of features (quadruple, human interactome network properties, and sequence similarity). There is script for each feature type, combining all converted data into a single file that is ready to pass as input to the classifier. After the data preprocessing stage has been completed successfully, the classifier model will be executed to output the result associated with the job ID.

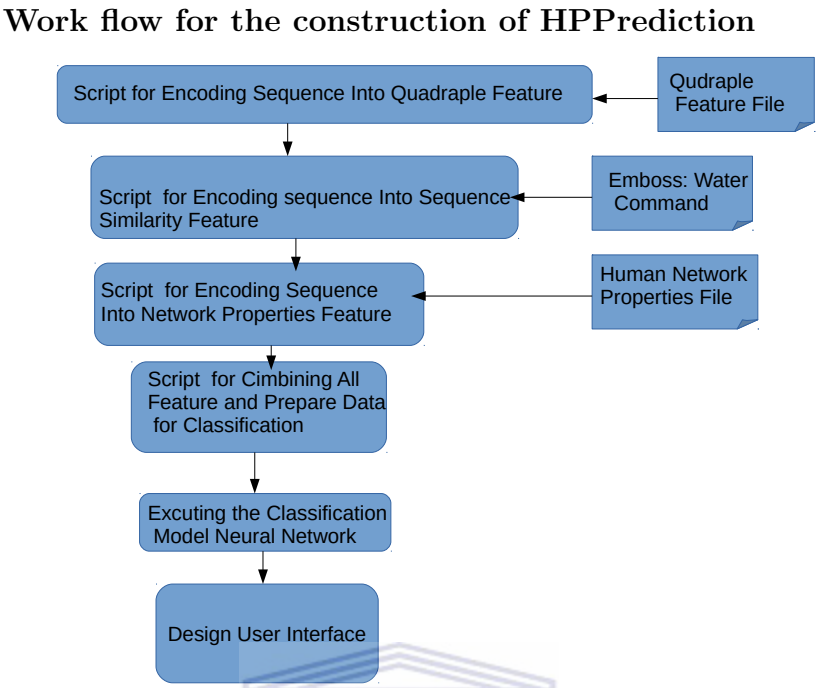


Figure 4.1: Work flow for the construction of HPPrediction web server. This diagram illus- trates the data parsing and binary classification model. It includes a web based user interface.

4.4 Description of the web server

4.4.1 Home page

The home page provide the user with a brief introduction to the web server.

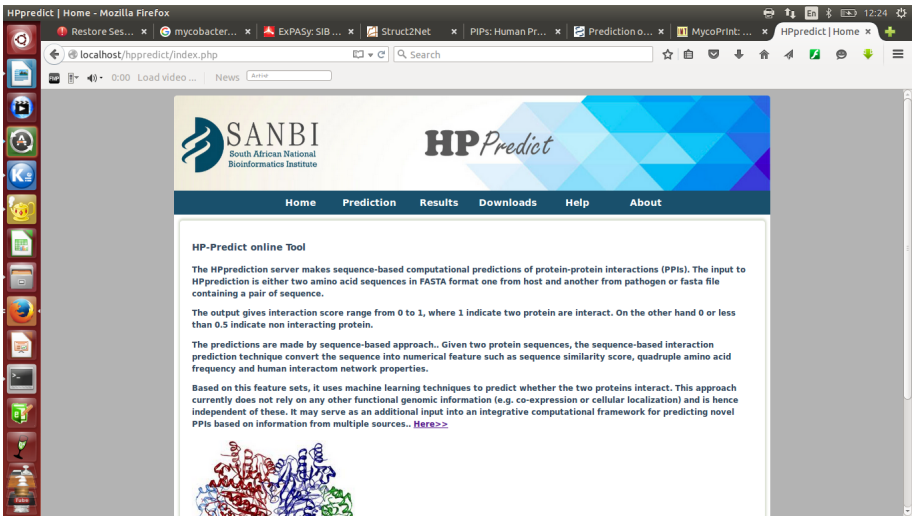


Figure 4.2: Host-pathogen. Input data includes a pair of sequences in FASTA format

4.4.2 Host-pathogen prediction

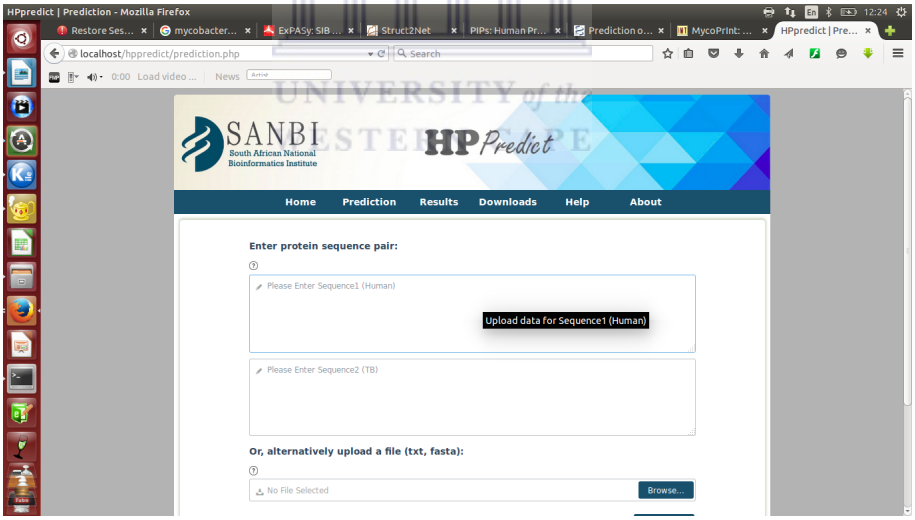


Figure 4.3:

The prediction page serves as the main engine of the predictor. The user submits a pair of protein sequences in FASTA format, Alternatively the prediction page provides a file uploading option. The file must contain a pair of amino acids from both host and pathogen. After pressing the submit button the unique job ID is generated. Using this job ID a user

can receive the result on result page.

>P9WQN5(*Mycobacterium tuberculosis*)
MGESERSEAFGIPRDSPLSSGDAAELEQLRREAAVLREQLENAVGSHAPTRSARDIHQLE
ARIDSLAARNSKLMETLKEARQQLLALREEVDRLGQPPSGYGVLLATHDDDTVDVFTSGR
KMRLTCSPNIDAASLKKGQTVRLNEALTVVEAGTFEAVGEISTLREILADGHRALVVGHA
DEERVVWLADPLIAEDLPDGLPEALNDDTRPRKLRPGDSLLVDTKAGYAFERIPKAEVED
LVLEEVPDVSYADIGGLSRQIEQIRDAVELPFLHKELYREYSLRPPKGVLLYGPPGCGKT
LIAKAVANSLAKKMAEVRGDDAHEAKSYFLNIKGPELLNKFVGETERHIRLIFQRAREKA
SEGTPVIVFFDEMDSIFRTRGTGVSSDVETTVVPQLLSEIDGVEGLENVIVIGASNREDM
IDPAILRPGRLDVKIKIERPDAEAAQDIYSKYLTEFLPVHADDLAEFDGDRSACIKAMIE
KVVDRMYAEIDDNRFLEVTYANGDKEVMYFKDFNSGAMIQNVVDRAKKNAIKSVLET-
GQP GLRIQHLLDSIVDEFAENEDLPNTTNPDDWARISGKKGERIVYIRTLVTGKSS-
SASRAID TESNLGQYL

>P11940(HUMAN)
MNPSAPSYPMASLYVGDLHPDVTEAMLYEKFSPAGPILSIRVCRDMITRRSLGYAYVNFQ
QPADAERALDTMNFDVIKGKPVRIMWSQRDPSLRKSGVGNIFIKNLDKSIDNKALY-
DTFS AFGNILSCKVVCDENGSKGYGFVHFETQEAAERAIEKMNGMLLNDRKVFV-
GRFKSRKERE AELGARAKEFTNVYIKNFGEDMDDERLKDLFGKFGPALSVKVMT-
DESGKSKGFGFVSFER HEDAQKAVDEMNGKELNGKQIYVGRAQKKVERQTELKRK-
FEQMKQDRITRYQGVNLYVKN LDDGIDDERLRKEFSPFGTITSAKVMMEGGRSKGFGFVCF-
SSPEEATKAVTEMNGRIVAT KPLYVALAQRKEERQAHLTNQYMQRMASVRAVPN-
PVINPYQPAPPSGYFMAAIPQTQNRA AYYPPSQIAQLRPSPRWTAQGARPHPFQN-
MPGAIRPAAPRPPFSTMRPASSQVPRVMSTQ RVANTSTQTMGPRPAAAAAAATPAVRTVPQYKY
PQQHLNAQPQVTMQQPAVHVQ GQEPLTASMLASAPPQEQKQMLGERLFPLIQAMH-
PTLAGKITGMLLEIDNSELLHMLESP ESLRSKVDEAVAVLQAHQAKEAAQKAVNSAT-
GVPTV

Alternatively the prediction page provides a file uploading option. The file must contain a pair of amino acids from both host and pathogen. After pressing the submit button the unique job ID is generated. Using this job ID a user can receive the result on result page.

4.4.3 Result Page

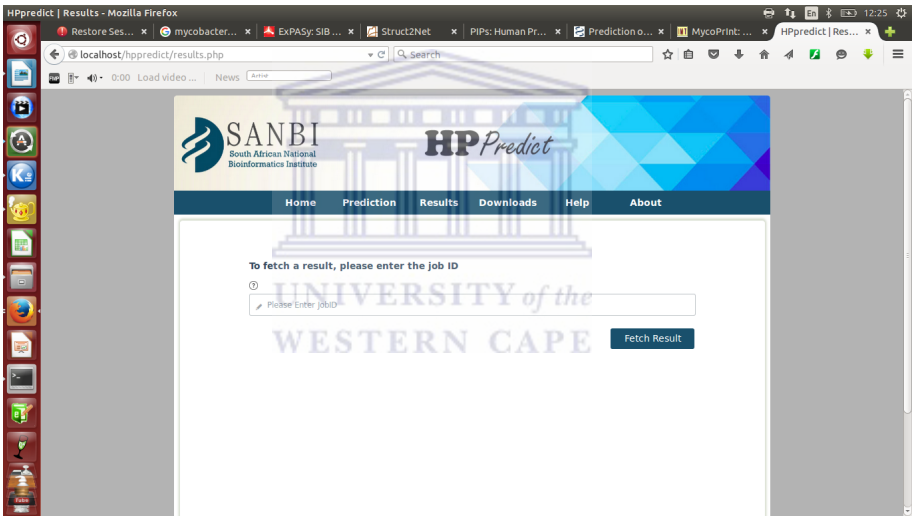


Figure 4.4:

After the data submission processes has been completed successfully, the engine will do all the calculation such as converting amino acid sequences to numerical data and calculating a prediction using the model. However, all these processes take place in the back-end system. The output is a prediction score that ranges from 0 to 1. Score from 0.6 to 1 means that two proteins are interacting, but the strength of interaction depends on how close the score is to 1.

4.4.4 Download Page

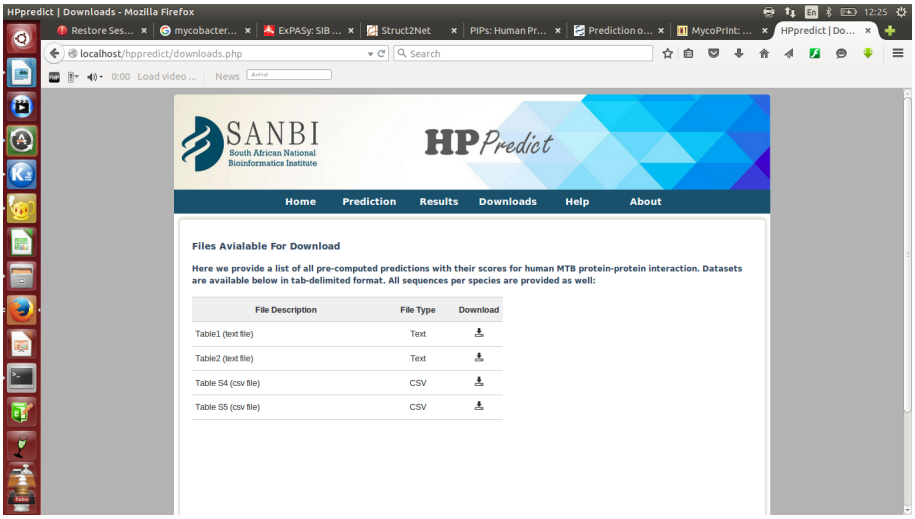


Figure 4.5:

The download page provides all the supplementary data that was used for developing these tools.

4.5 Conclusions

HPPredict is a convenient tool for the identification of potential protein-protein interactions between a host and pathogen, which may be vital in exploring drug targets for infectious diseases. This web server can also be used to construct the human-*M-tuberculosis* interactome network for a novel protein whose function is unknown. In general, one protein may interact with at least several partners including upstream and downstream regulators. These are useful guidelines for further experimental validation of the signaling network around any given protein. However, HPPredict currently provide models for human-*M-tuberculosis* PPI. Additional models must be generated for other infectious disease such as malaria. Another feature to be included in subsequent releases will be the addition of functional annotations for the predicted PPI so that the user can interpret their results. Finally, network analysis and visualization, are important for further elucidation

of the result. HPPredict provides a new type of tool to facilitate the prediction of direct or indirect protein partners and guides scientists to persue new experimental directions. The HPPredict server is available as a public web service http://sanbi/HPPrediction/.

Chapter 5

Conclusions and recommendations

Numerous human diseases are caused by bacterial infections. Our lack of understanding of the intimate relation between the pathogen and its host complicates the development of therapies. Protein-protein interactions are key players in every cell function, both within and between organisms, at every level of cellular function. Comprehensively identifying these interactions are essential to understand the mechanisms by which pathogens evade the hosts immune system. Past experimental and computational research largely focused on identifying interactions within single organisms. However, the elucidation of inter-species PPIs offers an alternative avenue for the design of novel chemotherapeutics. Prediction of critical PPIs in pathogens and host-pathogen systems will allow the design of several inhibitors at a given time. Thus, characterizing the interspecies interactome on a systematic level has only been a recent focus. High-throughput experimental techniques are being adapted to identify the interactions of both organisms at the same time. However, there is still no single cost-effective and highly accurate experimental technique to identify interactions on a large scale. As was the case for intra-species protein interactomes, computational methods could be utilized to inform and accelerate experimental endeavors. This thesis contributes towards identifying an interspecies interactome, specifically between human and *Mycobacterium tuberculosis*. A mechanine learning perspective was adopted throughout this thesis. The task of predicting PPIs were formulated in a binary classification framework, where each possible protein pair falls into one of two

classes, the interacting protein pairs (positive class) and the non-interacting protein pairs (negative class). The classifiers were trained in a supervised setting. In developing these predictors, several data and methodology-related challenges were handled. Firstly, Chapter 2 describes the first supervised model. A feed forward neural network classifier was employed to learn to distinguish interacting proteins from non-interacting pairs. One challenge to building such a system is identifying biological information that can serve as predictive features. Therefore, identifying information that is predictive in distinguishing interacting protein pairs from non-interacting ones is important. Thus, we developed a model that determines an optimal feature set that can be more representative for host-pathogen interaction prediction. This thesis demonstrated that the quadruple amino acid consecutive frequencies feature combined with human interactome properties plus pairwise sequence similarity score, can give optimal results on host-pathogen PPIs prediction. This result was validated in a comparison to published feature selection (Cui et al., 2012; Gomez et al., 2003; Taylor et al,. 2004) in human- HIV PPI data. Secondly, in Chapter 3 we utilized the optimal feature set obtained from Chapter 2 to develop a novel model for human-*Mycobacterium tuberculosis* PPI. In Chapter 3 we use tow intra-species PPI data in order to make inter-species prediction. To our knowledge there is no previous work that utilize tow intra-species data combined with machine learning to predict host-pathogen interactions. Therefore, we developed a model using human-*Bacillus anthracis* for which there is available inter- species PPI data. This approach was then extended to human-*M-tuberculosis* PPI prediction. Finally, in Chapter 4 we developed a web server to predict the likelihood of two human-*M-tuberculosis* proteins interacting based on the model developed in Chapter 3. There are several potential directions for future extensions of this research. Firstly, the web server can be expanded to incorporate functional information and visualization. Secondly, the server can be modified for predicting interactions in other Host-pathogen sys tems. Along with *M-tuberculosis*, there are many other clinically important pathogens for which computational models could shed light on their interaction with the human host. The binary classification setting I provide and most of the features I derive can be extended to predicting other host-pathogen PPIs. The limiting step will

be the availability of the labeled data.

numerous of human diseases are caused by bacterial infections. Our lack of understanding of the intimate relation between the pathogen and its host complicates the development of therapie. Protein-protein interactions are key players in every cell function, both within and between organisms, at every level of cellular function. Comprehensively identifying these interactions is essential towards discovering how cellular processes take place. Past experimental and computational research largely focused on identifying interactions within single organisms. However, the elucidation of inter-species PPIs offers an alternative avenue for the design of novel chemotherapeutics. Prediction of critical PPIs in pathogens and host-pathogen systems will allow the design of several inhibitors at a given time. Thus, characterizing the interspecies interactome on a systematic level has only been a recent focus. High-throughput experimental techniques have being adapted to handle the interactions of both organisms at the same time. However, there is still no single cost-effective and highly accurate experimental technique to identify interactions on a large scale. As was the case for intra-species protein interactome, computational methods could be utilized to inform and accelerate experimental endeavors.

This thesis contribute to identifying an interspecies interactome, specifically between human and *Mycobacterium tuberculosis*. Throughout the thesis, I employed a machine learning perspective. The task of predicting PPIs were formulated in a binary classification framework, where each possible protein pair falls into one of two classes, the interacting protein pairs (positive class) and the non-interacting protein pairs (negative class). The classifiers were learnt in a supervised setting. In developing these predictors, several data and methodology-related challenges were handled. Firstly, Chapter 2 describes the first supervised model. A feed forward neural network classifier was employed to learn to distinguish interacting proteins from non-interacting pairs. One challenge to building such a system is identifying biological information that can serve as predictive features. Therefore, identifying information that is predictive in distinguishing interacting protein pairs from non-interacting ones is important. Thus, we developed a model that determines

an optimal feature set that can be more representative for host-pathogen interaction prediction. However, the feature selection part shows that the quadruple amino acide consecutive frequencies feature combined with human interactome properties plus pair-wise sequence similarity score, can give optimal results on host-pathogen PPIs prediction. This result was validated in a comparison to published feature selection (Cui et al., 2012; Gomez et al., 2003; Taylor et al,. 2004) in human-HIV PPI data. Secondly, in Chapter 3 we utilized the optimal feature set obtained from Chapter 2 to develop a novel model for human-*Mycobacterium tuberculosis* PPI. In Chapter 3 we use intra-species PPI data in order to make inter-species prediction. To our knowledge there is no previous work that utilize intra-species data combined with machine learning to predict host-pathogen interactions. Therefore, we first developed a model on different species, named proof of concept model where there is available inter-species PPI data. After approval of the concept, then we proceed to implement it on huamn-*M-tuberculosis* prediction. Finally, in Chapter 4 we developed a web server which use the model implmented in Chapter 3.

There are several potential directions for future extensions of this research. Firstly, incorporating additional features to improve the accuracy of the host-pathogen prediction task. Secondly, the web server can be expanded to incorporate functional prediction and visualization. The server can be modified for predicting interactions in other Host-pathogen systems. Along with *M-tuberculosis*, there are many other clinically important pathogens for which computational models could shed light on their interaction with the human host. The binary classification setting I provide and most of the features I derive can be extended to predicting other host-pathogen PPIs. The limiting step will be the availability of the labeled data.

Bibliography

Andreas Z, Luisa MP, Michele Q, Gabriele A, Manuela HC, and Gianni C. MINT: a Molecular INTeraction database. *FEBS Letters*. 513 ((2002)) 135-140.

Aloy P, and Russell RB. InterPreTS: protein Interaction Prediction through Tertiary Structure. *Bioinformatics*. 19 (2003) 161-162.

Shoemaker AS, and Anna RP. Deciphering ProteinProtein Interactions. Part I. Experimental Techniques and Databases. *PLoS Computational Biology* 3 (2007).

Berman HM, Westbrook J, Feng Z, Gilliland1 G, Bhat TN, Weissig H, Shindyalov IN, and Bourne1 PE. The Protein Data Bank. *Nucl. Acids Res* 28 (2000) 235-242.

Bishop CM. *Pattern Recognition and Machine Learning* Springer Science + Business Media, LLC, New York (2006).

BenHur A, and Noble WS. Choosing negative examples for the prediction of protein-protein interactions.*BMC Bioinformatics* 7 (2006) 1-6.

BenHur A, Noble WS. Kernel methods for predicting protein-protein interactions. *Bioinformatics* 21 (Suppl.1) (2005) i38-i46.

Ben-Hur A, Ong CS, Sonnenburg S, Schölkopf B, Rätsch G. Support Vector Machines and Kernels for Computational Biology. *PLoS Comput Biol* 4 (2008). // Bock JR, and Gough DA .Predicting proteinprotein interactions from primary structure. *Bioinformatics* 17 (2001) 455-460.

Bradford JR, and Westhead DR. Improved prediction of proteinprotein binding sites using a support vector machines approach. *Bioinformatics* 21 (2005) 1487-1494.

Becker KG, Barnes KC, Bright TJ, and Wang AS. The Genetic Association Database. *Nature Genetics* 36 (2004) 431-432.

Beibarth T, and Speed TP. GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *BIOINFORMATICS* 20 (2004) 1464-1465.

Breiman L. Random forests. Machine Learning. 45 (2001) 5-32.

Becker KG, Barnes KC, Bright TJ, and Wang AS. The Genetic Association Database. *Nature Genetics* 36 (2004) 431-432.

Bulusu K.C., Tym J.E., Coker E.A., Schierz A.C., Al-Lazikani B. canSAR: updated cancer research and drug discovery knowledgebase. Nucleic Acids Res. 2014;42:D1040-D1047.

Cui G, Fang C and Han K. Prediction of protein-protein interactions between viruses and human by an SVM model. *BMC Bioinformatics* 13 (2012).

Chen D and Stow D . The Effect of Training Strategies on Supervised Classification at Different Spatial Resolutions. *Photogrammetric Engineering and*

Remote Sensing 68 (2002) 1155-1161.

Chen X, and Liu M. Prediction of proteinprotein interactions using random decision forest framework. *Bioinformatics* 21 (2005) 4394-4400.

Cristianini N. and ShaweTaylor J. *An Introduction to Support Vector Machines*. Cambridge University Press, Cambridge, UK. (2000).

Chinnasamy A, Mittal A, and Sung W. Probabilistic prediction of protein-protein interactions from the protein sequences. *Comput Bio Med* 36 (2006) 1143-1154.

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

Dyer MD, Neff C, Dufford M, and Rivera CG, Shattuck D, Riera J. B, Murali T. M, and Sobral B. W. The Human-Bacterial Pathogen Protein Interaction Networks of *Bacillus anthracis Francisella tularensis* and *Yersinia pestis*. *PLoS ONE* 5 (2010).

Deng M, Zhang K, Mehta S, Chen T, Sun F. Prediction of protein function using protein-protein interaction data. *Proc IEEE Comput Soc Bioinform Conf* 1 (2002) 197-206.

Davis FP, Barkan DT, Eswar N, McKerrow JH, and Sali A. Host-pathogen protein interactions predicted by comparative modeling *Protein Science* 16 (2007) 2585-2596.

Dyer MD, Murali TM, and Sobral BW. The Landscape of Human Proteins

Interacting with Viruses and Other Pathogens. *PLoS Pathog* 4 (2008).

Dyer MD, Murali TM, and Sobral BW. Computational prediction of host-pathogen protein-protein interactions. *Bioinformatics* 23 (2007) 159-166.

Dyer MD, Murali TM, and Sobral BW. Supervised learning and ptrediction of ysical interactions between human and HIV proteins. *Infection, Genetics and Evolution* 11 (2011) 917-923.

DongMel C, and Douglas S. The Effect of Training Strategies on Supervised Classification at Different Spatial Resolutions. *Photogrammetric Engineering and Remote Sensing* 68 (2002) 1155-1161.

Davis FP, Barkan DT, Eswar N, Mckerrow JH, and Sali A. Host-pathogen protein interactions predicted by omparative modeling. *Protein Science* 16 (2007) 2585-2596.

Dohkan S, Koike A, and Takagi T. Improving the performance of an SVM-based method for predicting protein-protein interactions. *In Silico Biol* 6 (2006) 515-529.

Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA. Protein interaction maps for complete genomes based on gene fusion events. Nature 4 (1999) 86-90.

Enright AJ, and Ouzounis CA. Functional associations of proteins in entire genomes by means of exhaustive detection of gene fusions. *Genome Biology* 2 (2001).

Enright AJ, Dongen SV, and Ouzounis CA. An efficient algorithm for large-

scale detection of protein families. *Nucleic Acids Research* 30 (2002) 1575-1584.

Eisenberg MAP, Marcotte EM, Thompson MJ, Eisenberg D, and Yeates TO. Assigning protein functions by comparative genome analysis:Protein phylogenetic profiles. *Proc. Natl. Acad. Sci. USA* 96 (1999) 4285-4288.

Fields S, Song O. A novel genetic system to detect protein-protein interactions. *Nature* 340 (1989) 245-246.

Finn RD, Miller BL, Clements J, Bateman A. iPfam: a database of protein family and domain interactions found in the Protein Data Bank. *Nucleic Acids Res* 42 (2014) D364-D373.

Gomez SM, Noble WS, and Rzhetsk A. Learning to predict proteinprotein interactions from protein sequences. *Bioinformatics* 19 (2003) 1875-1881.

Guo Y, Yu l, Wen Z, and Li M. Using support vector machine combined with auto covariance to predict proteinprotein interactions from protein sequences. *Nucleic Acids Research* 36 (2008) 3025-3030.

Guy E, Mallampalli A. Managing TB in the 21st century: existing and novel drug therapies. *Ther. Adv. Respir. Dis.* 2 (2008) 401-408.

Garcia-Garcia J, Schleker S, Klein-Seetharaman J, and Oliva B. BIPS: BIANA Interolog Prediction Server. A tool for protein-protein interaction inference . *leic Acids Research* 40 (2012) W147-W151.

Gomez SM, Noble WS, Rzhetsky A: Learning to predict protein-protein in-

teractions from protein sequences. *Bioinformatics* 19 (2003)1875-1881.

Taylor WR: The classification of amino acid conservation. *J Theor Biol* 119 (1986) 205-218.

Henning H, Luisa M. P, Chris L, Sugath M, Samuel K, Sandra O, Martin V, Bernd R, Peter R, Alfonso V, Hanah M, John A, Amos B, Gianni C, David S and Rolf A. IntAct: an open source molecular interaction database. *Nucleic Acids Research* 32 (2004) D452-D455.

Huang L, Bosch I, Hofmann W, Sodroski J, and Pardee AB. Tat Protein Induces Human Immunodeficiency Virus Type 1 (HIV-1) Coreceptors and Promotes Infection with both Macrophage-Tropic and T-Lymphotropic HIV-1 Strains. *Journal of Virology*, 72 (1998) 8952-8960.

Huang DW, Sherman BT, and Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. *Nature Protoc*4 (2009) 44-57.

Huynen M, Snel1 B, Warren L, and Bork P. Predicting Protein Function by Genomic Context: Quantitative Evaluation and Qualitative Inferences. *Genome Research* 10 (2000) 1204-1210.

Huo T, Liu W, Guo Y, Yang C, Lin J, and Rao Z. Prediction of host-pathogen protein interactions between *Mycobacterium tuberculosis* and Homo sapiens using sequence motifs. *BMC Bioinformatics* 16 (2015).

Holm L, and Rosenström P. Dali server: conservation mapping in 3D. *Nucl. Acids Res.* 38 (2010) W545-W549.

Ioannis X, Lukasz S, Xiaoqun JD, Patrick H, Sul Min. K and David E. DIP

the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Research* 30 (2002) 303-305.

Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan N. Y, Chung S, Emili A, Snyder M, Greenblatt J. F, and Gerstein M. A Bayesian Networks Approach for Predicting Protein-Protein Interactions from Genomic Data. *Science* 302 (2003) 449-453.

Jansen R, and Gerstein M. Analyzing protein function on a genomic scale: the importance of gold-standard positives and negatives for network prediction. *Curr Opin Microbiol* 7 (2004) 535-545.

Krishnadev O, Bisht S, and Srinivasan N. Prediction of Protein-Protein Interactions Between Human Host and Two Mycobacterial Organisms. *International Journal of Knowledge Discovery in Bioinformatics* 1 (2010) 1-13.

Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, Li J, Pu S, Datta N, Tikuisis AP, Punna T, Peregrn-Alvarez JM, Shales M, Zhang X, Davey M, Robinson MD, Paccanaro A, Bray JE, Sheung A, Beattie B, Richards DP, Canadien V, Lalev A, Mena F, Wong P, Starostine A, Canete MM, Vlasblom J, Wu S, Orsi C, Collins SR, Chandran S, Haw R, Rilstone JJ, Gandi K, Thompson NJ, Musso G, St Onge P, Ghanny S, Lam MH, Butland G, Altaf-Ul AM, Kanaya S, Shilatifard A, O?Shea E, Weissman JS, Ingles CJ, Hughes TR, Parkinson J, Gerstein M, Wodak SJ, Emili A, and Greenblatt JF. Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature* 440 (2006) 637-643.

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

Krishnadev O, and Srinivasan N. Prediction of protein-protein interactions between human host and a pathogen and its application to three pathogenic bacteria. *Int J Biol Macromol* 48 (2011) 613-619

LoBue P. Extensively drug-resistant tuberculosis. *Current Opinion in Infectious Diseases* 22 (2009) 167-173.

Lee SA, Chan C, Tsai CH, Lai JM, Wang FS, Kao CY, and Huang CYF. Ortholog-based protein-protein interaction prediction and its application to inter-species interactions. *Bmc Bioinformatics* 9 (2008).

Liu CH, Li KC, and Yuan S. Human Protein-Protein Interaction Prediction by A Novel Sequence-Based Coevolution Method: Coevolutionary Divergence. *Bioinformatics* 29 (2013) 92-98.

Marcotte EM, Pellegrini M, Ng H, Rice D. W, Yeates TO, and Eisenberg D. Detecting Protein Function and ProteinProtein Interactions from Genome Sequences. *Science* 285 (1999) 751-753.

Mitchell TM. *Machine Learning*. McGraw-Hill, New York, 1997.

McCulloch JL, and Pitts W. logical calculus of ideas immanen in nervous activity. *Bulletin of Mathematical Biophysics* 5 (1943) 115-133.

Lu L. J, Xia Y, Paccanaro A, Yu H, and Gerstein M. Assessing the limits of genomic data integration for predicting protein networks. *Genome Research* 15 (2005) 945-953.

Lee K., Sung M.-K., Kim J., Kim K., Byun J., Paik H., Kim B., Huh W.-K., Ideker T. Proteome-wide remodeling of protein location and function by stress. Proc. Natl. Acad. Sci. U.S.A. 2014;111:E3157-E3166.

Martin S, Roe D, and Faulon JL. Predicting proteinprotein interactions using signature products. *Bioinformatics* 21 (2005) 218-226.

Matthews LR, Vaglio P, Reboul J, Ge H, Davis BP, Garrels J, Vincent S, and Vidal M. Identification of Potential Interaction Networks Using Sequence-Based Searches for Conserved ProteinProtein Interactions or (Interologs). *Genome Research* 11 (2001) 2120-2126.

Mogensen TH, Paludan SR, Kilian M, and Astergaard L. Live Streptococcus pneumoniae Haemophilus influenzae and Neisseria meningitidis activate the inflammatory response through Toll-like receptors 2, 4, and 9 in species- specific patterns. *Journal of Leukocyte Biology* 80 (2006) 267-277.

Magrane M, Consortium U. UniProt Knowledgebase: a hub of integrated protein data. Database. *The Journal of Biological Databases and Curation* (2011).

Mazandu GK, Mulder NJ (2011) Generation and analysis of large-scale data-driven Mycobacterium tuberculosis functional networks for drug target identification. Adv Bioinformatics 2011: Article ID 801478.

Najafabadi HS, and Salavati R. Sequence-based prediction of protein-protein interactions by means of codon usage. *Genome Biology* 9 (2008).

Ogmen U, Keskin O, Aytuna AS, Nussinov R, and Gursoy A. PRISM: protein interactions by structural matching. *Nucleic Acids Res.* 33 (2005) W331-W336.

Ooi SL, Pan X, Peyser BD, Ye P, Meluh PB, Yuan DS, Irizarry RA, Bader

JS, Spencer FA, and Boeke JD. Global synthetic-lethality analysis and yeast functional profiling. *Trends Genet.* 22 (2006) 56-63.

Okuda S, Kawashima S, Goto S, and Kanehisa M .Conservation of Gene Co-Regulation between Two Prokaryotes: *Bacillus subtilis* and *Escherichia coli.Genome Inform* 16 (2005) 116-124.

Overbeek R, Fonstein M, D'Souza M, Pusch GD, and Maltsev N. The use of gene clusters to infer functional coupling.*Proceeding of National Academy of Sciences of the Unided State of America* . 96 (1999) 2896-2901.

Piter S, Alamgir M, Green JR, Dumontier M, Dehne F, and Golshani A. Computational methods for predicting protein-protein interactions. *Adv Biochem E* 110 (2008) 247-267.

Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, and Yeates TO. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proc Natl Acad Sci* 96 (1999) 4285-4288.

Philipp P, Stefan K, Matthias O, Barbara B, Irmtraud DK, Goar F, Corinna M, Pekka M, Volker S, Hans WM, Andreas R, and Dmitrij F. The MIPS mammalian protein-protein interaction database. *Bioinformatics* 6 (2005) 832-834.

Pagel P , Wong P, and Frishman D. A Domain Interaction Map Based on Phylogenetic Profiling. *J. Mol. Biol* 344 (2004) 1331-1346.

Planas-Iglesias J, Marin-Lopez MA, Bonet J, Garcia-Garcia J, and Oliva B. iLoops: a protein-protein interaction prediction server based on structural features. *Bioinformatics Advance Access* 29 (2013) 2360-2362.

Qi Y, Bar-Joseph Z, and Klein-Seetharaman J. Evaluation of different biological data and computational classification methods for use in protein interaction prediction. *Proteins* 63 (2006) 490500.

Rashid M, Ramasamy S, and Raghava GP. A Simple Approach for Predicting Protein-Protein Interactions. *Current Protein and Peptide Science* 11 (2010) 589-600.

Rapanoel HA, Mazandu GK, and Mulder NJ. Predicting and Analyzing Interactions between *Mycobacterium tuberculosis* and Its Human Host. *PLoS One* 8 (2013).

Rao VS, Srinivas K, Sujini GN, and Kumar GN. Protein-protein interaction detection: methods and analysis. *International journal of proteomics* 2014 (2014).

Ranea JAG, Yeats C, Grant A, and Orengo CA. Predicting Protein Function with Hierarchical Phylogenetic Profiles: The Gene3D Phylo-Tuner Method Applied to Eukaryotic Genomes. *PLoS Computational Biology* 3 (2007) 2366-2378

Skrabanek L, Saini HK, Bader GD, and Enright AJ. Computational Prediction of Protein-Protein Interactions. *Molecular Biotechnol* 38 (2008) 1-17.

Shoemaker BA, and Panchenko AR. Deciphering ProteinProtein Interactions.Part II. Computational Methods to Predict Protein and Domain Interaction Partners. *PLoS Computational Bi* 3 (2007) e43.

Sun j, Xu J,Liu Z, Liu Q, Zhao A, Shi T, and Li Y. Refined phylogenetic profiles method for predicting proteinprotein interactions. *Bioinformatics* 21 (2005) 3409-3415

Shannon P, Markiel A, Ozier O, Baliga N, Wang J. T, Ramage D, Amin N, Schwikowski B, and Ideker T. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks . *Genome Res* 13 (2003) 24982504.

Shen J, Zhang J, Luo X, Zhu W, Yu K, Chen K, Li Y, and Jiang H. Predicting protein-protein interactions based only on sequences information. *PNAS* 104 (2007) 43374341.

Soong TT, Wrzeszczynski KO, Rost B. Physical protein-protein interactions predicted from microarrays. *Bioinformatics* 24 (2008) 2608-2614.

Skrabanek L, Harpreet K. S, Bader G D, Enright A. J. Computational Prediction of Protein-Protein Interactions. *Mol Biotechnol* 38 (2008) 381-17.

Lin T. W, Wu J. W, and Chang D. T. Combining Phylogenetic Profiling-Based and Machine Learning-Based Techniques to Predict Functional RelatedProteins. *PLoS ONE* 8 (2013) e75940.

Levy E.D., Landry C.R., Michnick S.W. How perfect can protein interactomes be. Sci. Signal. 2009;2:pe11.
Simonsen M, Maetschke S, and Ragan M. A. Automatic Selection of Reference Taxa for Protein-Protein Interaction Prediction with Phylogenetic Profiling. *Bioinformatics* 28 (2012) 51-57.

Shen J, Zhang J, Luo X, Zhu W, Yu K, Chen K, Li Y, and Jiang H. Predicting proteinprotein interactions based only on sequences information. *PNAS* 104 (2007) 4337-4341.

Sato T, Yamanishi Y, Kanehisa M, and Toh H. Prediction of Protein-Protein

Interactions Based on Real-Valued Phylogenetic Profiles Using Partial Correlation Coefficient. *GIW* (2004) page 122.

Tyagi N, Krishnadev O, and Srinivasan N. Prediction of protein-protein interactions between Helicobacter pylori and a human host. *Mol. BioSyst* 5 (2009) 1630-1635.

Tastan O, Qi Y, Carbonell GJ, and Klein-Seetharaman J. Prediction of interactions between HIV-1 and human proteins by information integration. *Pac Symp Biocomput.* (2009) 516-527.

Tamames J, Casari G, Ouzounis C, Valencia A. Conserved Clusters of Functionally Related Genes in Two Bacterial Genomes. *Journal of Molecular Evolution* 73 (1997) 44-66.

Vidal M, Cusick ME, and Albert-Lszl Barabsi A. Interactome networks and human disease. *Cell* 144 (2011) 986-998.

Vidal M., Cusick M.E., Barabsi A.-L. Interactome networks and human disease. Cell 2011;144:986-998.
Vapnik VN. Support-vector networks. *Machine Learning* 20 (1995) 273-297.

Wattam AR , Abraham D, Dalay O, Disz TL, Driscoll T, Gabbard JL, Gillespie JJ, Gough R, Hix D, Kenyon R, Machi D, Mao C, Nordberg EK, Olson R, Overbeek R, Pusch GD, Shukla M, Schulman J, Stevens RL, Sullivan DE, Vonstein V, Warren A, Will R, Wilson MJ, Yoo HS, Zhang C, Zhang Y, and Sobral BW. .PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Res* 42 (2014) D581-591.

WHO http://www.who. int/tb/publications /global_report/gtbr12_main.pdf

Wuchty S. Computational prediction of host-parasite protein interactions between *plasmodium falciparum* and human. *PloS one* 6 (2011) e26960.

Werbos PJ. *Beyond regression: New tools for prediction and analysis in the behavioral sciences*. PhD thesis, Harvard University, Cambridge, MA. (1974)

Wattam AR, Abraham D, Dalay O, Disz TL, Driscoll T, Gabbard JL, Gillespie JJ, Gough R, Hix D, Kenyon R, Machi D, Mao C, Nordberg EK, Olson R, Overbeek R, Pusch GD, Shukla M, Schulman J, Stevens RL, Sullivan DE, Vonstein V, Warren A, Will R, Wilson MJ, Yoo HS, Zhang C, Zhang Y, Sobral BW . 2014. PATRIC, the bacterial bioinformatics database and analysis resource. Nucleic Acids Res. 42:D581D591. 10.1093/nar/gkt1099.

Wiwatwattana N., Kumar A. Organelle DB: a cross-species database of protein localization and function. Nucleic Acids Res. 2005;33:D598-D604.

Yaveroglu ON, and Can T. Predicting Protein-Protein Interactions from Protein Sequences Using Phylogenetic Profiles. *World Academy of Science, Engineering and Technolog* 56 (2009).

Zahiri J, Bozorgmehr JH, and Nejad AM. Computational Prediction of ProteinProtein Interaction Networks: Algorithms and Resources . *Current Genomics* 14 (2013) 397-414.

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

Zhou H, Jin J, Zhang H, Yi B, Wozniak M, and Wong L. IntPath–an inte-

grated pathway gene relationship database for model organisms and important pathogens. *BMC Syst Biol* 6 (2012).

Appendices

Appendix A

Supplementary material

A.1 Supplementary material for Chapter 2

Quadruple Feature Set

Sample of four consecutive amino acid frequencies, which we used as quadruple feature set.

('I', 'I', 'I', 'I'), ('I', 'I', 'I', 'F'), ('I', 'I', 'I', 'H'), ('I', 'I', 'I', 'D'), ('I', 'I', 'I', 'Q'), ('I', 'I', 'I', 'A'), ('I', 'I', 'F', 'I'), ('I', 'I', 'F', 'F'), ('I', 'I', 'F', 'H'), ('I', 'I', 'F', 'D'), ('I', 'I', 'F', 'Q'), ('I', 'I', 'F', 'A'), ('I', 'I', 'H', 'I'), ('I', 'I', 'H', 'F'), ('I', 'I', 'H', 'H'), ('I', 'I', 'H', 'D'), ('I', 'I', 'H', 'Q'), ('I', 'I', 'H', 'A'), ('I', 'I', 'D', 'I'), ('I', 'I', 'D', 'F'), ('I', 'I', 'D', 'H'), ('I', 'I', 'D', 'D'), ('I', 'I', 'D', 'Q'), ('I', 'I', 'D', 'A'), ('I', 'I', 'Q', 'I'), ('I', 'I', 'Q', 'F'), ('I', 'I', 'Q', 'H'), ('I', 'I', 'Q', 'D'), ('I', 'I', 'Q', 'Q'), ('I', 'I', 'Q', 'A'), ('I', 'I', 'A', 'I'), ('I', 'I', 'A', 'F'), ('I', 'I', 'A', 'H'), ('I', 'I', 'A', 'D'), ('I', 'I', 'A', 'Q'), ('I', 'I', 'A', 'A'), ('I', 'F', 'I', 'I'), ('I', 'F', 'I', 'F'), ('I', 'F', 'I', 'H'), ('I', 'F', 'I', 'D'), ('I', 'F', 'I', 'Q'), ('I', 'F', 'I', 'A'), ('I', 'F', 'F', 'I'), ('I', 'F', 'F', 'F'), ('I', 'F', 'F', 'H'), ('I', 'F', 'F', 'D'), ('I', 'F', ' F', 'Q'), ('I', 'F', 'F', 'A'), ('I', 'F', 'H', 'I'), ('I', 'F', 'H', 'F'), ('I', 'F', 'H', 'H'), ('I', 'F', 'H', 'D'), ('I', 'F', 'H', 'Q'), ('I', 'F', 'H', 'A'), ('I', 'F', 'D', 'I'), ('I', 'F', 'D', 'F'), ('I', 'F', 'D', 'H'), ('I', 'F', 'D', 'D'), ('I', 'F', 'D', 'Q'), ('I', 'F', 'D', 'A'), ('I', 'F', 'Q', 'I'), ('I', 'F', 'Q', 'F'), ('I', 'F', 'Q', 'H'), ('I', 'F', 'Q', 'D'), ('I', 'F', 'Q', 'Q'), ('I', 'F', 'Q', 'A'), ('I', 'F', 'A', 'I'),

('I', 'F', 'A', 'F'), ('I', 'F', 'A', 'H'), ('I', 'F', 'A', 'D'), ('I', 'F', 'A', 'Q'), ('I', 'F', 'A', 'A'), ('I', 'H', 'I', 'I'), ('I', 'H', 'I', 'F'), ('I', 'H', 'I', 'H'), ('I', 'H', 'I', 'D'), ('I', 'H', 'I', 'Q'), ('I', 'H', 'I', 'A'), ('I', 'H', 'F', 'I'), ('I', 'H', 'F', 'F'), ('I', 'H', 'F', 'H'), ('I', 'H', 'F', 'D'), ('I', 'H', 'F', 'Q'), ('I', 'H', 'F', 'A'), ('I', 'H', 'H', 'I'), ('I', 'H', 'H', 'F'), ('I', 'H', 'H', 'H'), ('I', 'H', 'H', 'D'), ('I', 'H', 'H', 'Q'), ('I', 'H', 'H', 'A'), ('I', 'H', 'D', 'I'), ('I', 'H', 'D', 'F'), ('I', 'H', 'D', 'H'), (' I', 'H', 'D', 'D'), ('I', 'H', 'D', 'Q'), ('I', 'H', 'D', 'A'), ('I', 'H', 'Q', 'I'), ('I', 'H', 'Q', 'F'), ('I', 'H', 'Q', 'H'), ('I', 'H', 'Q', 'D'), ('I', 'H', 'Q', 'Q'), ('I', 'H', 'Q', 'A'), ('I', 'H', 'A', 'I'), ('I', 'H', 'A', 'F'), ('I', 'H', 'A', 'H'), ('I', 'H', 'A', 'D'), ('I', 'H', 'A', 'Q'), ('I', 'H', 'A', 'A'), ('I', 'D', 'I', 'I'), ('I', 'D', 'I', 'F'), ('I', 'D', 'I', 'H'), ('I', 'D', 'I', 'D'), ('I', 'D', 'I', 'Q'), ('I', 'D', 'I', 'A'), ('I', 'D', 'F', 'I'), ('I', 'D', 'F', 'F'), ('I', 'D', 'F', 'H'), ('I', 'D', 'F', 'D'), ('I', 'D', 'F', 'Q'), ('I', 'D', 'F', 'A'), ('I', 'D', 'H', 'I'), ('I', 'D', 'H', 'F'), ('I', 'D', 'H', 'H'), ('I', 'D', 'H', 'D'), ('I', 'D', 'H', 'Q'), ('I', 'D', 'H', 'A'), ('I', 'D', 'D', 'I'), ('I', 'D', 'D', 'F'), ('I', 'D', 'D', 'H'), ('I', 'D', 'D', 'D'), ('I', 'D', 'D', 'Q'), ('I', 'D', 'D', 'A'), ('I', 'D', 'Q', 'I'), ('I', 'D', 'Q', 'F'), ('I', 'D', 'Q', 'H'), ('I', 'D', 'Q', 'D'), ('I', 'D', 'Q', 'Q'), ('I', 'D', 'Q', 'A'), ('I', 'D', 'A', 'I'), ('I', 'D', 'A' , 'F'), ('I', 'D', 'A', 'H'), ('I', 'D', 'A', 'D'), ('I', 'D', 'A', 'Q'), ('I', 'D', 'A', 'A'), ('I', 'Q', 'I', 'I'), ('I', 'Q', 'I', 'F'), ('I', 'Q', 'I', 'H'), ('I', 'Q', 'I', 'D'), ('I', 'Q', 'I', 'Q'), ('I', 'Q', 'I', 'A'), ('I', 'Q', 'F', 'I'), ('I', 'Q', 'F', 'F'), ('I', 'Q', 'F', 'H'), ('I', 'Q', 'F', 'D'), ('I', 'Q', 'F', 'Q'), ('I', 'Q', 'F', 'A'), ('I', 'Q', 'H', 'I'), ('I', 'Q', 'H', 'F'), ('I', 'Q', 'H', 'H'), ('I', 'Q', 'H', 'D'), ('I', 'Q', 'H', 'Q'), ('I', 'Q', 'H', 'A'), ('I', 'Q', 'D', 'I'), ('I', 'Q', 'D', 'F'), ('I', 'Q', 'D', 'H'), ('I', 'Q', 'D', 'D'), ('I', 'Q', 'D', 'Q'), ('I', 'Q', 'D', 'A'), ('I', 'Q', 'Q', 'I'), ('I', 'Q', 'Q', 'F'), ('I', 'Q', 'Q', 'H'), ('I', 'Q', 'Q', 'D'), ('I', 'Q', 'Q', 'Q'), ('I', 'Q', 'Q', 'A'), ('I', 'Q', 'A', 'I'), ('I', 'Q', 'A', 'F'), ('I', 'Q', 'A', 'H'), ('I', 'Q', 'A', 'D'), ('I', 'Q', 'A', 'Q'), ('I', 'Q', 'A', 'A'), ('I', 'A', 'I', 'I'), ('I', 'A', 'I', 'F')

Triple Feature Set

Sample of three consecutive amino acid frequencies, which we used as triple feature set.

('I', 'V', 'L'), ('I', 'V', 'M'), ('I', 'V', 'F'), ('I', 'V', 'Y'), ('I', 'V', 'W'), ('I', 'V', 'H'), ('I', 'V', 'K'), ('I', 'V', 'R'), ('I', 'V', 'D'), ('I', 'V', 'E'), ('I', 'V', 'Q'), ('I', 'V', 'N'), ('I',

'V', 'T'), ('I', 'V', 'P'), ('I', 'V', 'A'), ('I', 'V', 'C'), ('I', 'V', 'G'), ('I', 'V', 'S'), ('I', 'L', 'V'), ('I', 'L', 'M'), ('I', 'L', 'F'), ('I', 'L', 'Y'), ('I', 'L', 'W'), ('I', 'L', 'H'), ('I', 'L', 'K'), ('I', 'L', 'R'), ('I', 'L', 'D'), ('I', 'L', 'E'), ('I', 'L', 'Q'), ('I', 'L', 'N'), ('I', 'L', 'T'), ('I', 'L', 'P'), ('I', 'L', 'A'), ('I', 'L', 'C'), ('I', 'L', 'G'), ('I', 'L', 'S'), ('I', 'M', 'V'), ('I', 'M', 'L'), ('I', 'M', 'F'), ('I', 'M', 'Y'), ('I', 'M', 'W'), ('I', 'M', 'H'), ('I', 'M', 'K'), ('I', 'M', 'R'), ('I', 'M', 'D'), ('I', 'M', 'E'), ('I', 'M', 'Q'), ('I', 'M', 'N'), ('I', 'M', 'T'), ('I', 'M', 'P'), ('I', 'M', 'A'), ('I', 'M', 'C'), ('I', 'M', 'G'), ('I', 'M', 'S'), ('V', 'I', 'L'), ('V', 'I', 'M'), ('V', 'I', 'F'), ('V', 'I', 'Y'), ('V', 'I', 'W'), ('V', 'I', 'H'), ('V' , 'I', 'K'), ('V', 'I', 'R'), ('V', 'I', 'D'), ('V', 'I', 'E'), ('V', 'I', 'Q'), ('V', 'I', 'N'), ('V', 'I', 'T'), ('V', 'I', 'P'), ('V', 'I', 'A'), ('V', 'I', 'C'), ('V', 'I', 'G'), ('V', 'I', 'S'), ('V', 'L', 'I'), ('V', 'L', 'M'), ('V', 'L', 'F'), ('V', 'L', 'Y'), ('V', 'L', 'W'), ('V', 'L', 'H'), ('V', 'L', 'K'), ('V', 'L', 'R'), ('V', 'L', 'D'), ('V', 'L', 'E'), ('V', 'L', 'Q'), ('V', 'L', 'N'), ('V', 'L', 'T'), ('V', 'L', 'P'), ('V', 'L', 'A'), ('V', 'L', 'C'), ('V', 'L', 'G'), ('V', 'L', 'S'), ('V', 'M', 'I'), ('V', 'M', 'L'), ('V', 'M', 'F'), ('V', 'M', 'Y'), ('V', 'M', 'W'), ('V', 'M', 'H'), ('V', 'M', 'K'), ('V', 'M', 'R'), ('V', 'M', 'D'), ('V', 'M', 'E'), ('V', 'M', 'Q'), ('V', 'M', 'N'), ('V', 'M', 'T'), ('V', 'M', 'P'), ('V', 'M', 'A'), ('V', 'M', 'C'), ('V', 'M', 'G'), ('V', 'M', 'S'), ('L', 'I', 'V'), ('L', 'I', 'M'), ('L', 'I', 'F'), ('L', 'I', 'Y'), ('L', 'I', 'W'), ('L', 'I', 'H'), ('L', 'I', 'K'), ('L', 'I', 'R'), ('L', 'I', 'D'), ('L', 'I', 'E'), ('L', 'I', 'Q'), ('L', 'I', 'N'), ('L', ' I', 'T'), ('L', 'I', 'P'), ('L', 'I', 'A'), ('L', 'I', 'C'), ('L', 'I', 'G'), ('L', 'I', 'S'), ('L', 'V', 'I'), ('L', 'V', 'M'), ('L', 'V', 'F'), ('L', 'V', 'Y'), ('L', 'V', 'W'), ('L', 'V', 'H'), ('L', 'V', 'K'), ('L', 'V', 'R'), ('L', 'V', 'D'), ('L', 'V', 'E'), ('L', 'V', 'Q'), ('L', 'V', 'N'), ('L', 'V', 'T'), ('L', 'V', 'P'), ('L', 'V', 'A'), ('L', 'V', 'C'), ('L', 'V', 'G'), ('L', 'V', 'S'), ('L', 'M', 'I'), ('L', 'M', 'V'), ('L', 'M', 'F'), ('L', 'M', 'Y'), ('L', 'M', 'W'), ('L', 'M', 'H'), ('L', 'M', 'K'), ('L', 'M', 'R'), ('L', 'M', 'D'), ('L', 'M', 'E'), ('L', 'M', 'Q'), ('L', 'M', 'N'), ('L', 'M', 'T'), ('L', 'M', 'P'), ('L', 'M', 'A'), ('L', 'M', 'C'), ('L', 'M', 'G'), ('L', 'M', 'S'), ('M', 'I', 'V'), ('M', 'I', 'L'), ('M', 'I', 'F'), ('M', 'I', 'Y'), ('M', 'I', 'W'), ('M', 'I', 'H'), ('M', 'I', 'K'), ('M', 'I', 'R'), ('M', 'I', 'D'), ('M', 'I', 'E'), ('M', 'I', 'Q'), ('M', 'I', 'N'), ('M', 'I', 'T'), ('M', 'I', 'P'), ('M', 'I', 'A'), ('M', 'I', 'C'), ('M', 'I', 'G'), ('M', 'I', 'S'), ('M', 'V', 'I'), ('M', 'V', 'L'), ('M', 'V', 'F'), ('M', 'V', 'Y'), ('M', 'V', 'W'), ('M', 'V', 'H'), ('M', 'V', 'K'), ('M', 'V', 'R'), ('M', 'V', 'D'), ('M', 'V', 'E'), ('M',

'V', 'Q'), ('M', 'V', 'N'), ('M', 'V', 'T'), ('M', 'V', 'P'), ('M', 'V', 'A'), ('M', 'V', 'C'), ('M', 'V', 'G'), ('M', 'V', 'S'), ('M', 'L', 'I'), ('M', 'L', 'V'), ('M', 'L', 'F'), ('M', 'L', 'Y'), ('M', 'L', 'W'), ('M', 'L', 'H'), ('M', 'L', 'K'), ('M', 'L', 'R'), ('M', 'L', 'D'), ('M', 'L', 'E'), ('M', 'L', 'Q'), ('M', 'L', 'N'), ('M', 'L', 'T'), ('M', 'L', 'P'), ('M', 'L', 'A'), ('M', 'L', 'C'), ('M', 'L', 'G'), ('M', 'L', 'S')

Table A.1: Significantly enriched GO terms for human proteins predicted to interact with *Bacillus anthraces* based on artificial neural network using using DAVID database.

GO Term	Description	P-Value
GO:0051015	actin filament binding	2.752293578
GO:0042802	identical protein binding	8.0275229358
GO:0019899	enzyme binding	6.6513761468
GO:0008092	cytoskeletal protein binding	6.4220183486
GO:0043566	structure-specific DNA binding	2.9816513761
GO:0008134	transcription factor binding	6.1926605505
GO:0046983	protein dimerization activity	6.4220183486
GO:0016564	transcription repressor activity	4.3577981651
GO:0003690	double-stranded DNA binding	2.0642201835
GO:0003677	DNA binding	18.119266055
GO:0042803	protein homodimerization activity	4.128440367
GO:0030528	transcription regulator activity	12.6146788991
GO:0019900	kinase binding	2.752293578
GO:0043565	sequence-specific DNA binding	6.1926605505
GO:0048306	calcium-dependent protein binding	1.1467889908
GO:0016563	transcription activator activity	4.5871559633
GO:0043425	bHLH transcription factor binding	0.6880733945
GO:0003779	actin binding	3.8990825688

GO:0060589	nucleoside-triphosphatase regulator activity	4.5871559633
GO:0019901	protein kinase binding	2.2935779817
GO:0030695	GTPase regulator activity	4.3577981651
GO:0035258	steroid hormone receptor binding	1.1467889908
GO:0019903	protein phosphatase binding	1.1467889908
GO:0046982	protein heterodimerization activity	2.752293578
GO:0005083	small GTPase regulator activity	3.2110091743
GO:0003712	transcription cofactor activity	3.8990825688
GO:0051427	hormone receptor binding	1.6055045872
GO:0019902	phosphatase binding	1.1467889908
GO:0051082	unfolded protein binding	1.8348623853
GO:0050681	androgen receptor binding	0.9174311927
GO:0003714	transcription corepressor activity	2.0642201835
GO:0030742	GTP-dependent protein binding	0.6880733945
GO:0048365	Rac GTPase binding	0.6880733945
GO:0047485	protein N-terminus binding	1.376146789
GO:0015631	tubulin binding	1.6055045872
GO:0019904	protein domain specific binding	3.4403669725
GO:0003723	RNA binding	6.1926605505
GO:0035257	nuclear hormone receptor binding	1.376146789
GO:0043047	single-stranded telomeric DNA binding	0.4587155963
GO:0003697	single-stranded DNA binding	1.1467889908
GO:0019838	growth factor binding	1.6055045872
GO:0042162	telomeric DNA binding	0.6880733945
GO:0019865	immunoglobulin binding	0.6880733945
GO:0005096	GTPase activator activity	2.5229357798
GO:0003700	transcription factor activity	7.7981651376

GO:0010843	promoter binding	1.1467889908
GO:0005201	extracellular matrix structural constituent	1.376146789
GO:0032393	MHC class I receptor activity	0.6880733945
GO:0005086	ARF guanyl-nucleotide exchange factor activity	0.6880733945
GO:0003743	translation initiation factor activity	1.1467889908
GO:0042289	MHC class II protein binding	0.4587155963
GO:0030911	TPR domain binding	0.4587155963
GO:0005099	Ras GTPase activator activity	1.376146789
GO:0005085	guanyl-nucleotide exchange factor activity	1.8348623853
GO:0003702	RNA polymerase II transcription factor activity	2.5229357798
GO:0003713	transcription coactivator activity	2.2935779817

Table A.2: Significantly enriched GO terms for human proteins predicted to interact with *Bacillus anthraces* based on artificial neural network using using DAVID database.

GO Term	Description	P-Value
GO:0008066	glutamate receptor activity	3.6253776435
GO:0020037	heme binding	3.9274924471
GO:0046906	tetrapyrrole binding	3.9274924471
GO:0010851	cyclase regulator activity	1.5105740181
GO:0004672	protein kinase activity	8.4592145015
GO:0004674	protein serine/threonine kinase activity	6.6465256798
GO:0051119	sugar transmembrane transporter activity	1.8126888218

GO:0001640	adenylate cyclase inhibiting metabotropic glutamate receptor activity	1.2084592145
GO:0005355	glucose transmembrane transporter activity	1.5105740181
GO:0019825	oxygen binding	2.1148036254
GO:0005402	cation:sugar symporter activity	1.5105740181
GO:0005351	sugar:hydrogen symporter activity	1.5105740181
GO:0010853	cyclase activator activity	1.2084592145
GO:0030250	guanylate cyclase activator activity	1.2084592145
GO:0003677	DNA binding	20.8459214502
GO:0004970	ionotropic glutamate receptor activity	1.5105740181
GO:0015149	hexose transmembrane transporter activity	1.5105740181
GO:0030249	guanylate cyclase regulator activity	1.2084592145
GO:0009055	electron carrier activity	4.2296072508
GO:0005234	extracellular-glutamate-gated ion channel activity	1.5105740181
GO:0015145	monosaccharide transmembrane transporter activity	1.5105740181
GO:0015295	solute:hydrogen symporter activity	1.5105740181
GO:0070330	aromatase activity	1.2084592145
GO:0005070	SH3/SH2 adaptor activity	1.5105740181
GO:0005506	iron ion binding	3.9274924471
GO:0017076	purine nucleotide binding	15.4078549849
GO:0000166	nucleotide binding	17.5226586103
GO:0032555	purine ribonucleotide binding	14.501510574
GO:0032553	ribonucleotide binding	14.501510574

GO:0008395	steroid hydroxylase activity	0.9063444109
GO:0030554	adenyl nucleotide binding	12.6888217523
GO:0004373	glycogen (starch) synthase activity	0.6042296073
GO:0018685	alkane 1-monooxygenase activity	0.6042296073
GO:0008943	glyceraldehyde-3-phosphate dehydrogenase activity	0.6042296073
GO:0004357	glutamate-cysteine ligase activity	0.6042296073
GO:0001642	group III metabotropic glutamate receptor activity	0.6042296073
GO:0000774	adenyl-nucleotide exchange factor activity	0.6042296073
GO:0008067	metabotropic glutamate, GABA-B-like receptor activity	0.6042296073
GO:0060090	molecular adaptor activity	1.5105740181
GO:0001883	purine nucleoside binding	12.6888217523
GO:0032393	MHC class I receptor activity	0.9063444109
GO:0019992	diacylglycerol binding	1.5105740181
GO:0001882	nucleoside binding	12.6888217523
GO:0015631	tubulin binding	1.8126888218
GO:0005524	ATP binding	11.7824773414
GO:0008017	microtubule binding	1.5105740181
GO:0005230	extracellular ligand-gated ion channel activity	1.5105740181
GO:0032559	adenyl ribonucleotide binding	11.7824773414
GO:0032396	inhibitory MHC class I receptor activity	0.6042296073
GO:0051287	NAD or NADH binding	1.2084592145
GO:0050662	coenzyme binding	2.416918429
GO:0046983	protein dimerization activity	5.1359516616

GO:0008568	microtubule-severing ATPase activity	0.6042296073
GO:0051219	phosphoprotein binding	0.9063444109

A.2 Supplementary material for Chapter 3

A.2.1 Functional Enrichment Analysis

Table A.3: Significantly enriched GO terms for human proteins predicted to interact with *Mycobacterium tuberculosis* based on artificial neural network using using DAVID database.

GO Term	Description	P-Value
GO:0042611	MHC protein complex	3.23544921613908E-033
GO:0042613	MHC class II protein complex	4.82140931562631E-030
GO:0044459	plasma membrane part	1.35280381858508E-017
GO:0005615	extracellular space	7.23503205138012E-016
GO:0044421	extracellular region part	4.17180171700974E-013
GO:0005886	plasma membrane	7.11456314393665E-013
GO:0005887	integral to plasma membrane	1.10173418858245E-010
GO:0031226	intrinsic to plasma membrane	2.25524941885622E-010
GO:0005576	extracellular region	1.12807686190366E-007
GO:0009986	cell surface	3.88220988341176E-007
GO:0043020	NADPH oxidase complex	0.000002909
GO:0009897	external side of plasma membrane	9.90036867574453E-006
GO:0042612	MHC class I protein complex	2.31643760969542E-005
GO:0031224	intrinsic to membrane	0.0001789811
GO:0016021	integral to membrane	0.000253924
GO:0045121	membrane raft	0.0004491051

GO:0000267	cell fraction	0.0004504223
GO:0046696	lipopolysaccharide receptor complex	0.0015379475
GO:0005625	soluble fraction	0.002040066
GO:0042825	TAP complex	0.0031765206
GO:0042824	MHC class I peptide loading complex	0.0053559954
GO:0031982	vesicle	0.0103813822
GO:0043005	neuron projection	0.0113540174
GO:0005578	proteinaceous extracellular matrix	0.0200357283
GO:0043514	interleukin-12 complex	0.0250339938
GO:0045177	apical part of cell	0.0255253642
GO:0030141	secretory granulev	0.0264316574
GO:0005792	microsome	0.0304527589
GO:0031012	extracellular matrix	0.0305916012
GO:0055037	recycling endosome	0.0335509649
GO:0042598	vesicular fraction	0.0348446272
GO:0010008	endosome membrane	0.034982789
GO:0044440	endosomal part	0.034982789
GO:0030139	endocytic vesicle	0.0381731621
GO:0005768	endosome	0.0459514446
GO:0030870	Mre11 complex	0.0494450822
GO:0030425	dendrite	0.05536914
GO:0042995	cell projectionv	0.0567679446
GO:0043235	receptor complex	0.058834271
GO:0031410	cytoplasmic vesicle	0.0618097512
GO:0048471	perinuclear region of cytoplasm	0.0720397621
GO:0030670	phagocytic vesicle membrane	0.0732486703
GO:0005624	membrane fraction	0.0854260851

GO:0016324	apical plasma membrane	0.0872619927
GO:0005773	vacuole	0.0991265809

A.2.2 Cellular Compartment Analysis of Human Proteins Targeted by Predicted Host Pathogen PPIs.

Table A.4: Cellular compartment significantly enriched GO terms for human proteins predicted to interact with *Mycobacterium tuberculosis* based on artificial neural network using DAVID database.

GO:0042611	MHC protein complex	3.23544921613908E-033
GO:0042613	MHC class II protein complex	4.82140931562631E-030
GO:0044459	plasma membrane part	1.35280381858508E-017
GO:0005615	extracellular space	7.23503205138012E-016
GO:0044421	extracellular region part	4.17180171700974E-013
GO:0005886	plasma membrane	7.11456314393665E-013
GO:0005887	integral to plasma membrane	1.10173418858245E-010
GO:0031226	intrinsic to plasma membrane	2.25524941885622E-010
GO:0005576	extracellular region	1.12807686190366E-007
GO:0009986	cell surface	3.88220988341176E-007
GO:0043020	MHC protein complex	3.23544921613908E-033
GO:0042613	MHC class II protein complex	4.82140931562631E-030
GO:0044459	plasma membrane part	1.35280381858508E-017
GO:0005615	extracellular space	7.23503205138012E-016
GO:0044421	extracellular region part	4.17180171700974E-013
GO:0005886	plasma membrane	7.11456314393665E-013

GO:0005887	integral to plasma membrane	1.10173418858245E-010
GO:0031226	intrinsic to plasma membrane	2.25524941885622E-010
GO:0005576	extracellular region	1.12807686190366E-007
GO:0009986	cell surface	3.88220988341176E-007
GO:0043020	MHC protein complex	3.23544921613908E-033
GO:0042613	MHC class II protein complex	4.82140931562631E-030
GO:0044459	plasma membrane part	1.35280381858508E-017
GO:0005615	extracellular space	7.23503205138012E-016
GO:0044421	extracellular region part	4.17180171700974E-013
GO:0005886	plasma membrane	7.11456314393665E-013
GO:0005887	integral to plasma membrane	1.10173418858245E-010
GO:0031226	intrinsic to plasma membrane	2.25524941885622E-010
GO:0005576	extracellular region	1.12807686190366E-007
GO:0009986	cell surface	3.88220988341176E-007
GO:0043020	NADPH oxidase complex	0.000002909
GO:0009897	external side of plasma membrane	9.90036867574453E-006
GO:0042612	MHC class I protein complex	2.31643760969542E-005
GO:0031224	intrinsic to membrane	0.0001789811
GO:0016021	integral to membrane	0.000253924
GO:0045121	membrane raft	0.0004491051
GO:0000267	cell fraction	0.0004504223
GO:0046696	lipopolysaccharide receptor complex	0.0015379475
GO:0005625	soluble fraction	0.002040066
GO:0042825	TAP complex	0.0031765206
GO:0042824	MHC class I peptide loading complex	0.0053559954
GO:0031982	vesicle	0.0103813822
GO:0043005	neuron projection	0.0113540174

GO:0005578	proteinaceous extracellular matrix	0.0200357283
GO:0043514	interleukin-12 complex	0.0250339938
GO:0045177	apical part of cell	0.0255253642
GO:0030141	secretory granule	0.0264316574
GO:0005792	microsome	0.0304527589
GO:0031012	extracellular matrix	0.0305916012
GO:0055037	recycling endosome	0.0335509649
GO:0042598	vesicular fraction	0.0348446272
GO:0010008	endosome membrane	0.034982789
GO:0044440	endosomal part	0.034982789
GO:0030139	endocytic vesicle	0.0381731621
GO:0005768	endosome	0.0459514446
GO:0030870	Mre11 complex	0.0494450822
GO:0030425	dendrite	0.05536914
GO:0042995	cell projection	0.0567679446
GO:0043235	receptor complex	0.058834271
GO:0031410	cytoplasmic vesicle	0.0618097512
GO:0048471	perinuclear region of cytoplasm	0.0720397621
GO:0030670	phagocytic vesicle membrane	0.0732486703
GO:0005624	membrane fraction	0.0854260851
GO:0016324	apical plasma membrane	0.0872619927
GO:0005773	vacuole	0.0991265809
GO:0009897	external side of plasma membrane	9.90036867574453E-006
GO:0042612	MHC class I protein complex	2.31643760969542E-005
GO:0031224	intrinsic to membrane	0.0001789811
GO:0016021	integral to membrane	0.000253924
GO:0045121	membrane raft	0.0004491051
GO:0000267	cell fraction	0.0004504223

GO:0046696	lipopolysaccharide receptor complex	0.0015379475
GO:0005625	soluble fraction	0.002040066
GO:0042825	TAP complex	0.0031765206
GO:0042824	MHC class I peptide loading complex	0.0053559954
GO:0031982	vesicle	0.0103813822
GO:0043005	neuron projection	0.0113540174
GO:0005578	proteinaceous extracellular matrix	0.0200357283
GO:0043514	interleukin-12 complex	0.0250339938
GO:0045177	apical part of cell	0.0255253642
GO:0030141	secretory granule	0.0264316574
GO:0005792	microsome	0.0304527589
GO:0031012	extracellular matrix	0.0305916012
GO:0055037	recycling endosome	0.0335509649
GO:0042598	vesicular fraction	0.0348446272
GO:0010008	endosome membrane	0.034982789
GO:0044440	endosomal part	0.034982789
GO:0030139	endocytic vesicle	0.0381731621
GO:0005768	endosome	0.0459514446
GO:0030870	Mre11 complex	0.0494450822
GO:0030425	dendrite	0.05536914
GO:0042995	cell projection	0.0567679446
GO:0043235	receptor complex	0.058834271
GO:0031410	cytoplasmic vesicle	0.0618097512
GO:0048471	perinuclear region of cytoplasm	0.0720397621
GO:0030670	phagocytic vesicle membrane	0.0732486703
GO:0005624	membrane fraction	0.0854260851
GO:0016324	apical plasma membrane	0.0872619927

GO:0005773	vacuole	0.0991265809
GO:0009897	external side of plasma membrane	9.90036867574453E-006
GO:0042612	MHC class I protein complex	2.31643760969542E-005
GO:0031224	intrinsic to membrane	0.0001789811
GO:0016021	integral to membrane	0.000253924
GO:0045121	membrane raft	0.0004491051
GO:0000267	cell fraction	0.0004504223
GO:0046696	lipopolysaccharide receptor complex	0.0015379475
GO:0005625	soluble fraction	0.002040066
GO:0042825	TAP complex	0.0031765206
GO:0042824	MHC class I peptide loading complex	0.0053559954
GO:0031982	vesicle	0.0103813822
GO:0043005	neuron projection	0.0113540174
GO:0005578	proteinaceous extracellular matrix	0.0200357283
GO:0043514	interleukin-12 complex	0.0250339938
GO:0045177	apical part of cell	0.0255253642
GO:0030141	secretory granule	0.0264316574
GO:0005792	microsome	0.0304527589
GO:0031012	extracellular matrix	0.0305916012
GO:0055037	recycling endosome	0.0335509649
GO:0042598	vesicular fraction	0.0348446272
GO:0010008	endosome membrane	0.034982789
GO:0044440	endosomal part	0.034982789
GO:0030139	endocytic vesicle	0.0381731621
GO:0005768	endosome	0.0459514446
GO:0030870	Mre11 complex	0.0494450822
GO:0030425	dendrite	0.05536914

GO:0042995	cell projection	0.0567679446
GO:0043235	receptor complex	0.058834271
GO:0031410	cytoplasmic vesicle	0.0618097512
GO:0048471	perinuclear region of cytoplasm	0.0720397621
GO:0030670	phagocytic vesicle membrane	0.0732486703
GO:0005624	membrane fraction	0.0854260851
GO:0016324	apical plasma membrane	0.0872619927
GO:0005773	vacuole	0.0991265809

A.2.3 Pathway Enrichment Analysis

Table A.5: Significantly enriched pathways for human proteins involved in the predicted host-pathogen PPIs dataset.

Term	Description	P-value
hsa05330	Allograft rejection	1.43961320988873E-022
hsa04940	Type I diabetes mellitus	6.63451812361417E-021
hsa05332	Graft-versus-host disease	1.91008931630884E-018
hsa04620	Toll-like receptor signaling pathway	3.52723400581991E-016
hsa05320	Autoimmune thyroid disease	4.32686009511781E-016
hsa04060	Cytokine-cytokine receptor interaction	1.79254275419825E-015
hsa05310	Asthma	5.72234157506714E-013
hsa04672	Intestinal immune network for IgA production	2.31359584030868E-012
hsa04612	Antigen processing and presentation	4.06974087441091E-011
hsa05322	Systemic lupus erythematosus	4.88182455888323E-008

hsa05416	Viral myocarditis	6.1708141275611E-008
hsa04630	Jak-STAT signaling pathway	4.59215307449203E-007
hsa04640	Hematopoietic cell lineage	5.48180015452256E-007
hsa04514	Cell adhesion molecules (CAMs)	1.00650660616269E-005
hsa04650	Natural killer cell mediated cytotoxicity	0.000245965
hsa04621	NOD-like receptor signaling pathway	0.0005455903
hsa04623	Cytosolic DNA-sensing pathway	0.0016455838
hsa00980	Metabolism of xenobiotics by cytochrome	P450 0.002584611
hsa00982	Drug metabolism	0.0030556494
hsa04622	RIG-I-like receptor signaling pathway	0.0060044489
hsa05020	Prion diseases	0.0085270892
hsa04062	Chemokine signaling pathway	0.0116853081
hsa04660	T cell receptor signaling pathway	0.0127952358
hsa00590	Arachidonic acid metabolism	0.0411967332
hsa04210	Apoptosis	0.0522176682
hsa04670	Leukocyte transendothelial migration	0.0573589241
hsa04614	Renin-angiotensin system	0.0587534383
hsa04144	Endocytosis	0.0655169229