IN SILICO PREDICTION OF PROTEIN-PROTEIN INTERACTIONS BETWEEN *THEILERIA PARVA* AND THE *BOS TAURUS*

EVERLYN MUTHONI KAMAU

MASTER OF SCIENCE

(Molecular Biology and Bioinformatics)

JOMO KENYATTA UNIVERSITY OF AGRICULTURE AND TECHNOLOGY

2018

In silico prediction of protein-protein interactions between *Theileria parva* and the *Bos taurus*

Everlyn Muthoni Kamau

A thesis submitted in partial fulfillment for the degree of Master of Science in Molecular Biology and Bioinformatics in the Jomo Kenyatta University of Agriculture and Technology.

2018

DECLARATION

This thesis is my original work and has not been submitted for a degree in any other University

Signature…..... Date ….....

Everlyn [redacted] Kamau

This thesis has been submitted for examination with our approval as University supervisors.

Signature …..... Date …..

Dr. Steven Ger Nyanjom, PhD

JKUAT, Kenya

Signature …..... Date …..

Dr. Joseph Ng'ang'a, PhD

JKUAT, Kenya

Signature …..... Date …..

Dr. Mark Wamalwa, PhD

ILRI, Kenya

ACKNOWLEDGEMENT

Special thanks to my mother for her immense and endless support throughout my graduate education. I extend my gratitude to William Weir, Brian Shiels and Jane Kinniard, all of University of Glasgow for kindly providing various data sets used in this project. I am also thankful to Gordon Langsley and his team (Malak Haidar, Nadia Echebli, Mehdi Metheni) at Institut Cochin, France for allowing me to work with them during my stay in France, where I was fortunate to expand my knowledge and interests on host-pathogen systems biology. I am also grateful to my thesis advisors, Dr.s' Mark Wamalwa, Steven G. Nyanjom and Joseph Ng'ang'a for their willingness to work with me. Many thanks to Dr. Wamalwa for his help and patiently explaining concepts to me at the beginning of the project. Not to forget, Joyce Njuguna at BecA-ILRI who patiently taught me a lot on programming and for her immense help during the project. Special thanks to Alan Orth at ILRI for allowing me to carry out my work at the ILRI high performance computing server, and also for his constant assistance and guidance. I gratefully acknowledge the support provided to me by BecA-ILRI ABCF research fellowship scheme to conduct my project and I am thankful that my stay at ILRI also broadened my interests in science.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

ABBREVIATIONS AND ACRONYMS

PPI:	Protein-Protein Interactions
NF-kB:	Nuclear Factor kappa Beta
TNF:	Tumor Necrosis Factor
JNK:	c-Jun NH2-terminal kinase
IKK:	IkB Kinase
HSP:	Heat Shock Protein
DNA:	Deoxyribonucleic acid
STRING:	Search Tool for the Retrieval of Interacting Genes/Proteins
MINT:	Molecular Interaction database
BioGRID:	Biological General Repository for Interaction Datasets
MIPS:	Munich Information Center for Protein Sequences
mRNA:	messenger Ribonucleic acid
XIAP:	X-Linked Inhibitor Of Apoptosis
c-IAP:	Inhibitor of Apoptosis Protein 1
VIL1:	Villin 1
AKN3:	AT-Hook Transcription Factor 3
FAM82B:	Regulator of microtubule dynamics protein 1
CAPZB:	Capping Protein (Actin Filament) Muscle Z-Line, Beta
CAPZA:	Capping Protein (Actin Filament) Muscle Z-Line, Alpha
VEGF:	Vascular Endothelial Growth Factor
EGF:	Epidermal Growth Factor
FGF:	Fibroblast Growth Factor

PDGF: Platelet-derived Growth Factor

IRAK1: Interleukin-1 receptor-associated kinase 1

ATP5A1: ATP Synthase, H+ Transporting, Mitochondrial F1 Complex, Alpha Subunit 1

TUBB: Tubulin

PI-3K: Phosphatidylinositol-4,5-bisphosphate 3-kinase

FYN : Tyrosine-protein kinase Fyn

NEMO: NF-kappa-B essential modulator

FOS: cFos

JUN: cJun

PAK1: Protein (Cdc42/Rac) Activated Kinase 1

GO: Gene Ontology

PKB: Protein Kinase B

ABSTRACT

Theileria parva induces pathogenesis, characteristic of cancer cell transformation and associated with invasion, proliferation and altered gene expression of infected bovine host leukocytes. Protein interactions are important for biological functions that underlie processes essential to pathogenesis during infection and can be used to select potential therapeutic targets. Using information on conserved protein interactions in other organisms (interologs), protein interactions and orthologous relationships were predicted between *Theileria parva* and *Bos taurus* (the bovine mammalian host). Among the predicted interactions were *Theileria's* HSP90 and glutaredoxin-like protein, and bovine c-JUN, AKT1, Rac1, STAT3 and HIF1-α proteins, observed as hubs connecting the predicted interactions to protein interactions within host. Bovine proteins were enriched in pathways that reflect known phenotype of *Theileria* infection such as induction or inhibition of apoptosis signaling, metastasis and tissue invasion, IL-10 signaling, NF-κB/IKK activation, PI-3K pathway, TGF-β signaling, modulation of immune and inflammatory responses. Support vector machine classifiers trained with the predicted interactions identified known protein interactions with 86.22% accuracy, 84.72% precision, 89.88% sensitivity and 84.39% specificity measures. Predicted interactions provide insight into *Theileria*- and bovine-encoded interactions that contribute to infection, providing a candidate set for subsequent experimental studies with possible use for defining functional annotation to uncharacterized parasite proteins.

CHAPTER ONE

INTRODUCTION

1.1 Background information

Theileria parva is a protozoan parasite that infects T and B lymphocytes of domestic cattle and wild ruminants. *Theileria parva* are transmitted by ixodid ticks cyclo-propagatively and trans-stadially. *Theileria* sporozoites invade leukocytes, where they undergo schizogony and merogony developmental stages. *Theileria* merozoites are liberated from their host cells and invade red blood cells where they develop immediately into gamonts giving rise to the piroplasm stage (Bishop *et al*., 2004). There are no repeated cycles of red blood cell invasion. The pathogenic stages for *Theileria* are the transformed schizont-infected leukocytes and not the piroplasm-infected erythrocytes (Heussler *et al*., 2006). In the bovine host, the parasite develops within the lymphocyte into a schizont, transforming the target cell into a clonally proliferating lymphoblast. The parasite divides in synchrony with the host cells, and the parasitized cells invade tissues throughout the body (Figure 1), resulting in a severe and often rapidly fatal lympho-proliferative disease, East Coast Fever (ECF) (Shaw, 2003). In response to an unknown signal, merogony follows schizogony, producing the piroplasm blood stages, which are infective to engorging ticks. All parasite stages are haploid except for the short diploid zygote stage in the intestinal cells (Geysen *et al*., 1999). African buffalo is the natural reservoir of the parasite. Transmission of buffalo-derived *T. parva* to cattle results in a rapidly lethal disease,

but in many cases the parasites do not differentiate to the erythrocyte-infective stage and are not transmissible by ticks (Bishop *et al*., 2004).

Vaccination against ECF is based on an infection and treatment method consisting of inoculating cattle with a defined dose of live sporozoites and a simultaneous injection of a long-standing tetracycline to control the infection (Akoolo *et al*., 2008). This type of immunization requires mixture of parasite strains, providing long term immunity against homologous parasite strain but variable protection against heterologous strains. Recent studies on *T. parva* epitopes identification and immune response identified a number of Tp1 and Tp2 epitopes that are dominant targets of the CD8+ T-cell response in cattle (Graham *et al*., 2008; MacHugh *et al*., 2009). Protective immune response also develops in cattle, which recover from infection with *T. parva*, and this immunity is due to destruction of schizont-infected cells by cytotoxic T lymphocytes (Bishop *et al*., 2001; Pelle *et al*., 2011).

ECF is endemic in Sub-Saharan African and Asian countries, causing a serious economical losses concentrated on small-scale resource-poor households with an estimated 100% mortality especially in exotic breeds (Bishop *et al*., 2002). In Kenya, *T. parva* infection poses a significant threat to the livestock sector through the economic impact of the disease from cattle morbidity and mortality and production losses in all production systems, as well as from the costs of the measures taken to control ticks and the disease (Gachohi *et al*., 2012).

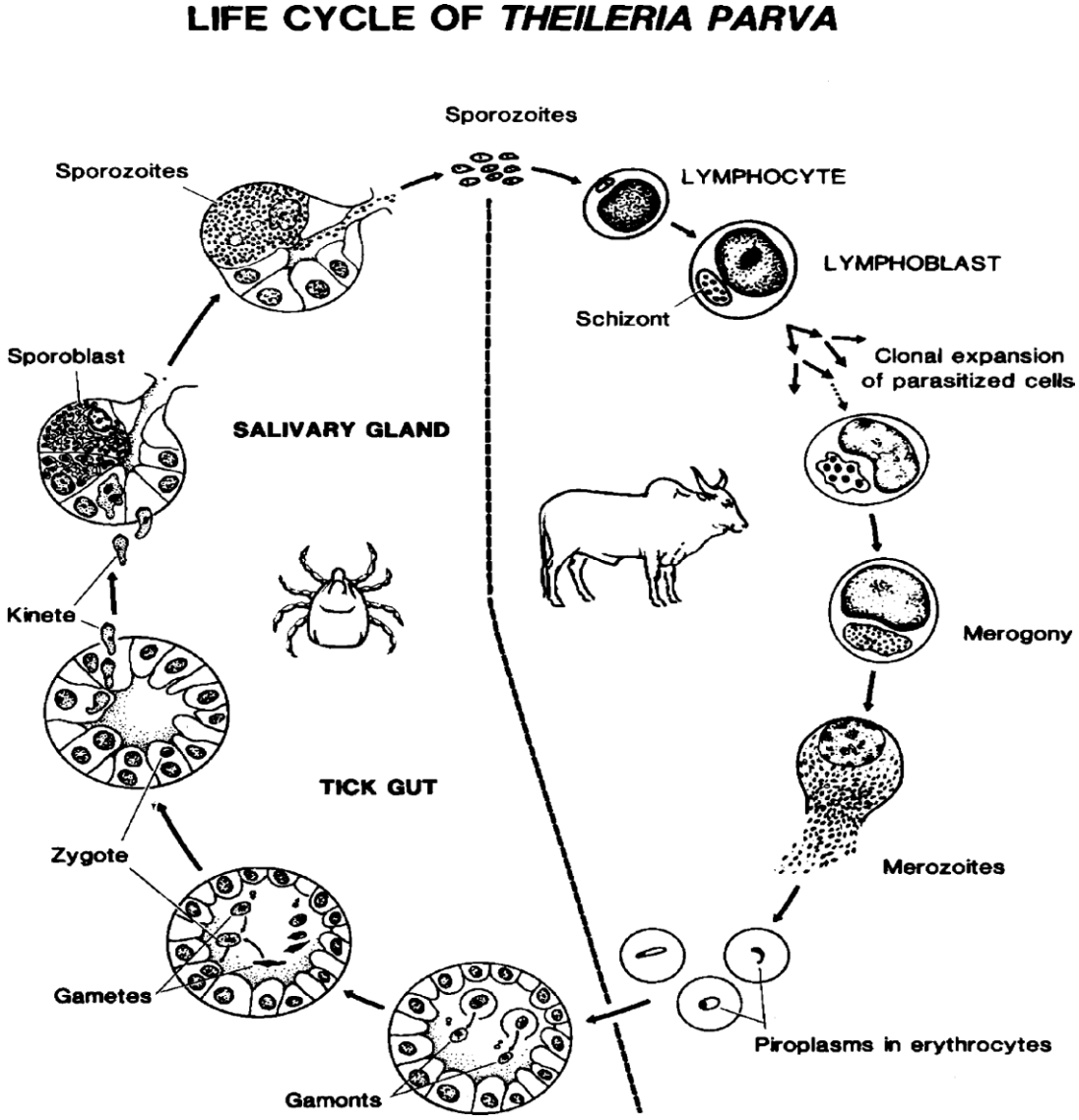

Figure 1.1: Life cycle of *Theileria parva* in cattle and in the ixodid tick vector *Rhipicephalus appendiculatus* (Bishop *et al*., 2002).

1.2 Intracellular protozoan parasites interaction with host cells

Parasitic protozoa adapt to their hosts through modulation or exploitation of cell death in order to facilitate parasite survival in a hostile environment. Very often, the pathogens use resources of the host cell, produce oxidative active substances which are exported into the host cell cytoplasm, interact with the host cell cytoskeleton and use host cell mitochondria as an additional source of nutrients (Luder *et al*., 2009). Mitochondrial apoptotic pathway may be modulated or exploited as part of the interactions between parasites and their hosts (Tato & Hunter, 2002; Schaumburg *et al*., 2006). Protozoa exert diverse effects on the host cell's fate including induction and inhibition of apoptotic cell death as well as simultaneously triggering both pro- and anti-apoptotic activities. Alternatively, induction of apoptosis during late stages of infection possibly facilitates dissemination and shedding of merozoites into the environment (Plattner & Soldati-Favre, 2008).

1.3 Problem Statement

Parasite-dependent regulation of host cell signalling pathways associated with *Theilerial* pathology results in East Coast Fever disease inflicting heavy economic losses to livestock-dependent livelihoods. ECF causes serious economical losses concentrated on small-scale resource-poor households, the morbidity rate is 100% among indigenous cattle, and about 50 million cattle at risk (with 10 million calves per annum) and the total yearly cost of the disease is estimated to be US $596

million. *Theileria* parasite has evolved to ensure its survival. However, the parasite molecules responsible for the initiation or regulation of the host cell transformation events are yet to be identified or fully validated, limiting drug or therapeutic targets against the pathogen, therefore limiting mechanisms of management of the disease. Pathogen-host interaction prediction is worthwhile to enlighten the infection mechanisms in the scarcity of experimentally verified interaction data.

1.4 Justification of the study

Computational prediction of interactions between host and pathogen proteins has previously been used to study human - *P. falciparum*, human-HCV, Salmonella-human, HPV-human and human-HIV systems, among others. Virus-host interactome screens for Influenza and Ebola were previously used to identify targetable host factors and guide antiviral drug development. Similarly, elucidating cellular processes and protein interaction networks during *Theileria parva* infection will help us understand better how pathogens infect their hosts and possibly understand disease processes at the protein level. Knowledge of probable interacting proteins and their functional classes will give a clearer picture of the important roles interactions may play in ECF pathogenesis. In addition, experimental methods are limited by low interaction coverage along with biases toward certain protein types and cellular localizations.

1.5 Hypothesis

Computational methods may not be used to predict inter-molecular interactions.

1.6 Objectives

1.6.1 Main objective

To predict *T. parva* and *B. taurus* protein interaction network (PIN) in ECF pathogenesis.

1.6.2 Specific objectives

i) To identify *T. parva* proteins that target and possibly remodel bovine host cell pathways.

ii) To predict inter-molecular interactions between *T. parva* and bovine host proteins.

iii) To identify key hubs (proteins involved in many interactions with other proteins) and bottlenecks (proteins central to many biological pathways).

CHAPTER TWO

LITERATURE REVIEW

2.1 *Theileria parva* subversion of bovine host cell regulatory pathways

Theileria sporozoites are introduced with the tick saliva as the tick takes a blood meal and rapidly invade leukocytes via yet unidentified receptors. Shortly after entry, the enveloping host cell membrane is eliminated and the parasite develops into a multi-nucleated macroschizont form, which resides freely in the host cell cytoplasm rather than in a parasitophorous vacuole (Brown *et al*., 1973). The free parasite then becomes permanently surrounded by an orderly array of polymerized host cell microtubules. This way, the parasite escapes lysosomal destruction and, from its cytoplasmic location, the schizont is perfectly poised to interfere with host cell signalling pathways that regulate host cell proliferation, apoptosis, proliferation, gene expression and survival (Shaw, 1997; Dobbelaere & Kuenzi, 2004). This lifestyle provides a more suitable for a fast exchange of information between the parasite and the host cell, because parasite molecules have to pass only a single membrane barrier to reach the signalling machinery within the host cell cytoplasm (Heussler *et al*., 2006). During cell division, *T. parva* parasites disseminate to both daughter cells by attachment to the mitotic spindle. These unique abilities result in clonal expansion of *Theileria*-infected host cells in a parasite-induced 'tumori-genesis' (Heussler *et al*., 1999). Other *T. parva*-associated phenotypic changes include production of a number of surface receptors, adhesion molecules and presentation of infection associated

antigens; alterations which could contribute to disease pathogenesis and subversion of the immune response (Dobbelaere & Heussler, 1999).

2.1.1 *T. parva* modulation of apoptosis

Theileria, among other obligate intracellular protozoans, rely on intact host cells to grow, propagate and differentiate (Plattner & Soldati-Fevre, 2008). They have evolved mechanisms to suppress or delay apoptosis in infected host cells, important prerequisites for sustained intracellular survival and development. They inhibit spontaneous host cell apoptosis or apoptosis induced by stress-related pro-apoptotic stimuli arguing for a selective advantage for those which are able to block the natural suicide programme (Dobbeleare *et al*., 1999; Ku¨enzi *et al*. 2003). Constitutive activation of the transcription factor NF-kB represents an important mechanism to prevent the parasite-transformed T host cells from apoptosis (Heussler *et al*., 1999; Tato & Hunter, 2002). Activation of NF-kB is accomplished by the recruitment and activation of the components of the IκB kinase (IKK) complex to the parasite surface, which then leads to the degradation of IκBα and consequently to the nuclear import of NF-kB (Schmuckli-Maurer *et al*., 2007; Heussler *et al*. 2002). NF-kB-dependent up-regulation of anti-apoptotic proteins also provides protection against TNF-induced apoptosis (Figure 2) (Heussler *et al*., 2001b; Heussler *et al*., 2006).

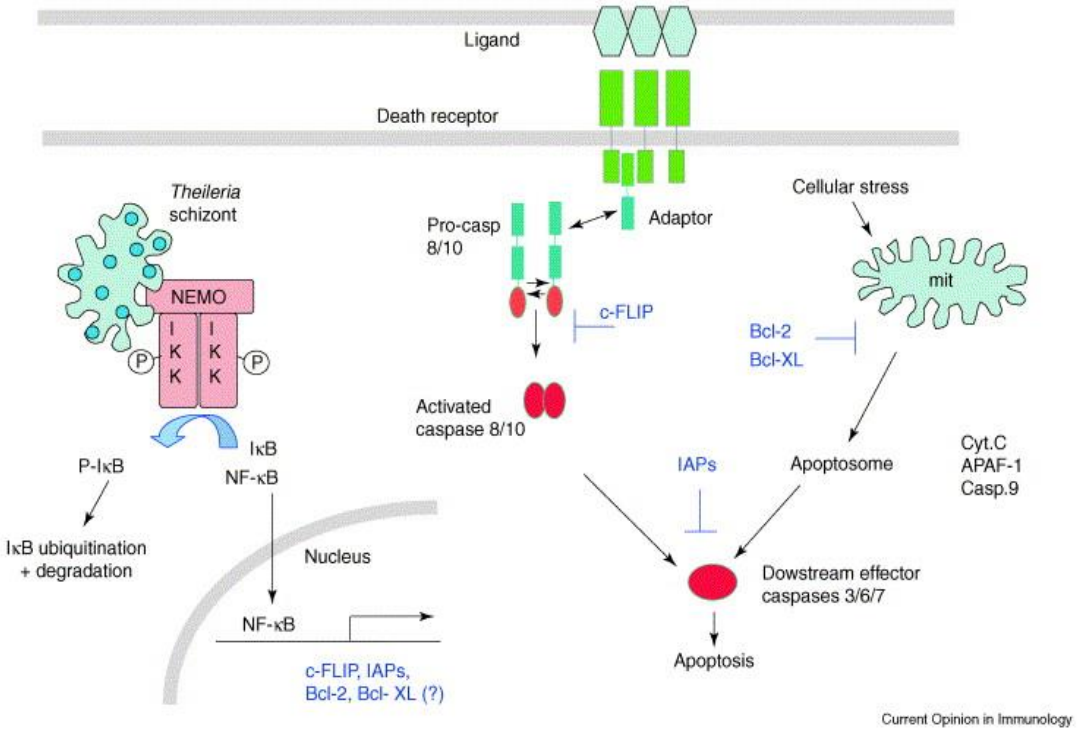

Figure 2.1 Model showing how Theileria-transformed cells are protected against apoptosis (Heussler *et al*., 2012)

T. parva employs XIAP and c-IAP 1 to render T cells resistant to Fas-induced apoptosis, since elimination of the parasite leads to the rapid decrease of these IAPs and to a concomitant increase in Fas-mediated apoptosis (Luder *et al*., 2009). Activation of the c-Jun N-terminal kinase (JNK) has also been implicated in the anti-apoptotic activity of *Theileria* (Lizundia *et al*. 2005; Lizundia *et al*. 2006). Increased transcription of the proto-oncogene *c-Myc* reduces – via the induction of anti-

apoptotic *Mcl-1* – the activation of caspase 9, the mitochondrial pathway of apoptosis in *Theileria*-infected B lymphocytes (Figure 2.2).

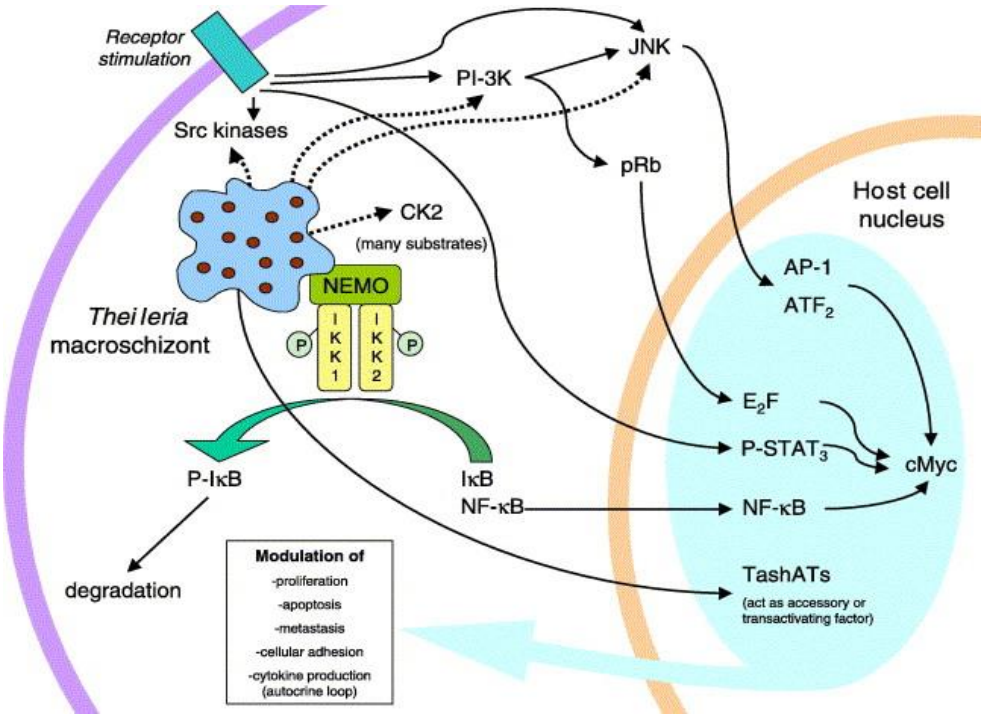

Figure 2.2: Pathways implicated in manipulation of *Theileria*-infected host cell phenotype (Dessauge *et al*. 2005a).

2.2 Protein-protein interactions

Molecular basis of cellular operations including regulation of metabolic pathways, immunologic recognition, DNA replication, transcription control, progression through the cell cycle, and protein synthesis are sustained largely by different

interactions among proteins. Many proteins participate in cellular processes as members of protein complexes of varying size. Correctly identifying the set of interacting proteins in an organism is useful for deciphering the molecular mechanisms underlying given biological functions and for assigning functions to unknown proteins based on their interacting partners (Qi *et al*., 2006). PPI network is presented as a graph, where nodes in the graph represent proteins and the edges that connect them correspond to interactions. Commonly, graph properties of the network, such as the degree or connection of nodes, the number and complexity of highly connected subgraphs, the shortest path length for indirectly connected nodes, alternative paths in the network and fragile key nodes are determined (von Mering *et al*., 2002).

Protein interactions can be classified into different types depending on their strength (permanent and transient), specificity, location of interacting partners within one or on two polypeptide chains, and the similarity between interacting subunits (homo- and hetero-oligomers) (Shoemaker & Pancheko, 2007a). Interactions between pathogen and host proteins allow pathogenic microorganisms to manipulate host mechanisms in order to use host capabilities and to escape from host immune responses (Dyer *et al*., 2010). With protein interactions, discovery of novel targets becomes easier since drugs discovered based on the protein interaction network may specifically modulate the disease-related pathways rather than simply inhibit or activate the functions of an individual target protein (Gonzalez & Kann, 2012). Globally conserved inter-species bacterial PPIs in host-pathogen interactomes were

used to derive novel drug targets in *C. pseudotuberculosis, C. diphtheriae, M. tuberculosis, C. ulcerans, Y. pestis, and E. coli* as targeted by Piper betel compounds (Barh *et al*., 2013).

Most of the previous studies concerning identification of protein interactions and their associated networks have been primarily focused on determining protein-protein interactions (PPIs) within a single organism (intra-species PPI prediction), while the prediction of PPIs between different organisms (inter-species PPI prediction) has been limited. Pathogen-host interaction prediction is worthwhile to enlighten the infection mechanisms in the scarcity of experimentally-verified interaction data. A number of databases and data resources store information on protein-protein interactions, molecular complexes and pathways from both computational and experimental techniques, including InterPreTS, STRING, GeneCensus, DIP, MINT, BioGRID and IntACT (Shoemaker & Pancheko, 2007b; Skrabanek *et al*., 2007).

2.3 Computational approaches for protein interactions prediction

Computational approaches utilize the structural, genomic, and biological context of proteins and genes to predict protein interaction networks and functional linkages between proteins (Marcotte *et al*., 1999; Yamanishi *et al*., 2004). Certain information and characteristics of protein have predictive value when integrated in combination and these include protein domains, similarity in mRNA expression profiles,

subcellular localization and possibility that proteins in a complex are bound by the same transcription factors (Zhang *et al*., 2004).

A few methods utilize genomics information to predict protein interactions e.g. conventional phylogenetic profiles including phylogenetic tree similarity; genomic proximity; gene co-localization and co-expression; gene fission / fusion; gene neighbourhood and transgenic distance (Jothi *et al*., 2006). Domain-based protein interaction prediction methods postulate that conservation of sequence properties such as domains, motifs, and signatures over the course of evolution may contribute to the interaction of proteins (Sprinzak & Margalit, 2001).

Machine learning techniques involve classifiers such as logistic regression (LR), random forests (RF), RF similarity-based nearest-neighbour (NN), decision trees (DT), naïve bayes (NB), bayesian networks and support vector machines (SVM) (Qi *et al*., 2006; Burger & Nimwegen, 2008). With machine learning techniques, interaction mining is also used to train learning systems to recognize correlated patterns within protein interaction pairs. Machine learning algorithms also utilize features and properties related to interface topology, solvent Accessible Surface Area (ASA) and hydrophobicity, or recognition of specific residue, geometric and conserved network motifs (Shen *et al*., 2007).

2.3.1 Support vector machines (SVM) for predicting protein interactions

SVM is a method that performs classification tasks by constructing hyperplanes in a multidimensional space that separates different class labels. Vectors in a training set S are projected to a higher-dimensional feature space. SVM supports both regression and classification tasks, and can handle multiple continuous and categorical variables. For categorical variables, a dummy variable is created with case values as 0 or 1 (Ben-Hur & Noble, 2005). A classification task involves separating data into training and testing sets. Each instance in the training set contains one target value (i.e. class labels) and several attributes (i.e. the features or observed variables). SVM produces a model, based on training data, which predicts the target values of the test data given only the test data attributes (Bradford & Westhead, 2005). Separation of data into training and testing sets is to assess the performance of prediction model on unseen data.

SVM provides the best generalization ability on unseen data compared with other classifiers for instance decision trees, neural networks (Bradford & Westhead, 2005). In constructing an SVM classifier, the support vectors are closest to the hyperplane and are located on the boundaries of the margin between two classes (Ben-Hur *et al*., 2008). SVM algorithms seek a function that defines a global hyperplane that would optimally separate the classes of training vectors with possible maximization of the margin between them. The margin corresponds to the distance between the points residing on the two edges of the hyperplane. The larger the margin, the better the generalization performance (Cristianini & Shawe-Taylor, 2000).

2.4 Ortholog based Protein-Protein Interaction prediction

Orthologs are proteins in different species that evolved from a common ancestor by speciation and have the same function. Conserved interactions between a pair of proteins, which have interacting homologs in another species, are referred to as "Interologs." The simple method of identifying interologs is as follows: Consider a template PPI pair (a, b) in a source species, find the homolog a in the host and the homolog b in the pathogen, conclude that (a, b) interact. If interacting proteins A and B in one organism have interacting orthologs A' and B' in another species, the pair of interactions A-B and A'-B' are called interologs (Yu *et al*., 2004). Protein-protein interactions can be transferred based on sequence similarities to homologs of both interacting partners (Lee *et al*., 2008). It is expected that the interactions between conserved orthologs, which are conserved genes and gene products in different species, will be conserved as in closely related species. The orthologs of co-evolving proteins also tend to interact, thereby making it possible to infer unknown interactions in other genomes (Matthews *et al*., 2001).

2.5 Features of protein interaction networks (PINs)

Computational methods for predicting PPIs exploit known protein and domain interactions, and information on sequence of proteins. Network topology measures can complement these data. Network topological features include hubs and bottlenecks. Bottlenecks correspond to the key connectors and dynamic components

of the interaction network (Yu *et al*., 2007). Hubs are proteins with high degrees and represent the most vulnerable points in a network. Between-ness measures the total number of non-redundant shortest paths going through a certain node or edge. Most of the shortest paths in a network go through the nodes with high between-ness, and they are the central points controlling the communication among other nodes in a network (Milo *et al*., 2002).

CHAPTER THREE

MATERIALS AND METHODS

3.1 Generating training datasets

Machine learning based methods, which formulate PPI prediction as a classification task use both interacting and non-interacting protein pairs as positive and negative classes, respectively.

3.1.1 Positive training dataset

Initial protein interactions between *Theileria parva* and *Bos taurus* were predicted using BIANA Interolog Prediction Server (BIPS) (Garcia-Garcia *et al*., 2010). BIANA integrates multiple sources of interaction data (HPRD, MINT, BioGrid, IntAct, MIPS). BIPS online interface was used to identify interacting proteins between *T. parva* and *B. taurus* (Hereford breed, NCBI reference genome annotation: 105) based on orthologous information and sequence homology filters as follows: 1e-10 e-value, 50% identity, 40% query sequence coverage, 80% template sequence coverage, 1e-10 joint e-value, and 30% joint identities. *T. parva* proteins previously identified as candidate manipulators of the infected leukocyte with potential to subvert host cell regulatory pathways (Graham *et al*., 2006; Shiels *et al*., 2006) were used to identify their putative interacting partners. These proteins are either secreted into host cell cytoplasm or located on parasite membrane. Sequences

for the identified interacting protein pairs were retrieved from UniProtKB using their unique ID's and BioPython parsers generating a training dataset of 5824 binary interactions.

3.1.2 Negative training dataset

A negative training dataset is essential to the reliability of a prediction model. However, the choice and lack of empirical non-interacting data is sometimes difficult since there are no empirically known non-interacting proteins. Artificial protein sequences were generated using an R script (Appendix 2) and randomly paired using a *Python* script to create a set of 5800 protein pairs. Protein pairs were paired using a *Python* script, based on recommendations for machine learning algorithms.

3.2 Generation of testing datasets

Protein interaction data containing 600 proteins was retrieved from HPRD Release 9 (Human Protein Reference Database, www.hprd.org) to form the positive training data. Protein sequences for the interacting pairs were then retrieved in batch in FASTA format from NCBI Genbank using BioPerl parsers. An equal number (600) of negative testing pairs were created as described above. Similarly, the two sets were concatenated to create an independent test dataset of 1200 protein pairs.

3.3 Reducing redundancy of protein sequences

All the sequences of the training and testing datasets were refined to reduce similarity between sequences using CD-HIT (Cluster Database at High Identity with Tolerance) (Li & Godzik, 2006). A redundancy threshold parameter at 25% was used resulting in a training set of 11472 protein pairs from the initial total of 11624 pairs.

3.4 Support vector machines (SVM) models

Prediction models were generated with SVM from the downloaded protein sequences. Each protein sequence in an interaction pair was represented by a vector space containing features of amino acids, and concatenating the vector spaces of two protein partners representing an interacting pair.

3.4.1 Protein features and vector encoding

Feature encoding transforms protein sequences of variable length to fixed length by capturing important information of protein sequence content in real number values. All the protein data was encoded using PROFEAT server (Rao *et al*., 2011). For each protein, an input vector of 1017 dimensions was generated based on four feature groups: amino acid composition, di-peptide composition, auto-correlation descriptors of physico-chemical properties and composition, transition and distribution properties. Autocorrelation descriptors are defined based on the distribution of amino

acid properties such as hydrophobicity, polarizability, normalized van-der-waals volume, polarity and net charge (Kawashima & Kanehisa, 2000, http://www.genome.jp.dbget-bin/aaindex.html).

The feature vectors were then labeled as '+1' for positive instance, '-1' for negative instance. Feature values in each vector were normalized and linearly scaled to the range of [-1,+1]. Lastly, the encoded datasets were indexed into SVM format and feature vectors for every pair of interacting proteins $\{D_A\}$ and $\{D_B\}$ were concatenated as $\{D_{AB,}\ y\}$ using *Python* scripts where y is the class label. For an interacting protein pair the two proteins were concatenated resulting to a vector of 2034 dimensions.

3.4.2 Model selection (parameter search)

To select a model and good parameters for the Gaussian radial basis function (RBF) kernel, C and γ, a coarse grid search approach was adopted within a limited range. log$C$ and logγ ranged from -3 to 12 and -3 to 15, respectively. Grid search approach tries various pairs of exponentially growing sequences of (C, γ) and the pair with the best cross-validation accuracy is selected. Grid (factorial) search for parameter selection was implemented in LIBSVM (Figure 3.1, search grid for parameter optimization). 5-fold cross-validation (CV) was used to identify a combination of values of cost C and kernel width γ parameters with best performance, minimal test error and good overall accuracy. Cross-validation randomly divides the dataset into 5

equal parts, four parts are used for training and the remaining fifth part is used as validation test set. The procedure is repeated 5 times such that each set is tested once.

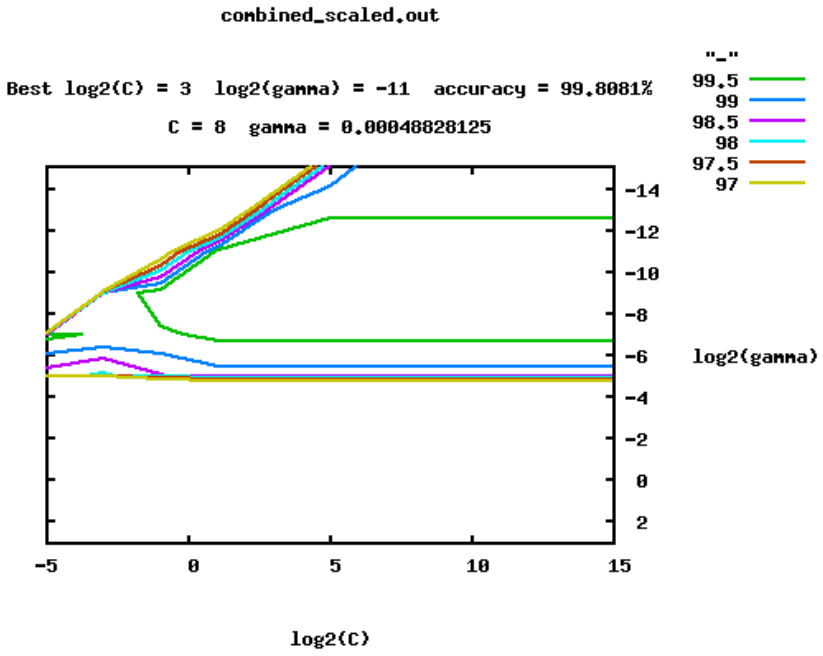

Figure 3.1: Contour plot of grid search results showing optimal values of *C* and *gamma* on training set data. 3-Dimension accuracy surface of 5-fold cross-validation on training set versus variations of C and gamma parameters.

3.4.3 Training an inductive SVM

SVMlight (Joachims, 2002; Joachims, 1999) was implemented to develop support vector machines classifiers and predict classes of testing dataset using '*svm_learn*'

and '*svm_classify*' algorithms/modules. Two classifiers based on linear and RBF kernels were used for the training dataset. RBF kernel non-linearly maps samples into a higher dimensional space, and particularly useful when the number of features is small compared to that of training instances. For the linear model, learning parameters were set as default in the *SVMlight* program, while *C* and γ were set at 8.0 and 0.0048 respectively for the RBF kernel model. The parameters set for RBF kernel model had been obtained from the grid search process at a 5-fold cross validation accuracy of 99%.

3.4.4 Performance measures

To assess performance of the inductive models these threshold dependent measures were calculated: sensitivity, specificity, accuracy, precision and Matthews correlation coefficient (MCC). *Precision* is the probability that an example predicted to be in class '+' truly belongs to this class, and *Recall* is the probability that an example belonging to class '+' is classified into this class. MCC considers both under- and over-predictions, where MCC=1 denotes a perfect prediction, MCC = 0 indicates a completely random assignment. Overall accuracy presents how well a classifier distinguishes true positives and true negatives, and 100% accuracy denotes a perfect prediction. A Receiver Operating Characteristic (ROC) curve was plotted for true positive rate “tpr” (sensitivity, y-axis) against false positive rate “fpr” (1-specificity, x-axis). A ROC curve shows the trade-off between sensitivity and specificity, the closer the curve follows the left-hand border and then the top border of the ROC

space, the more accurate the model, the closer the curve comes to the 45-degree diagonal of the ROC, the less the accurate the test; and the area under the curve (AUC) is a measure of accuracy. AUC of 1 represents a perfect model or ability to correctly classify data into two groups. 'ROCR' CRAN package in R to plot ROC curve was used to compute AUC and other quantitative performance measures.

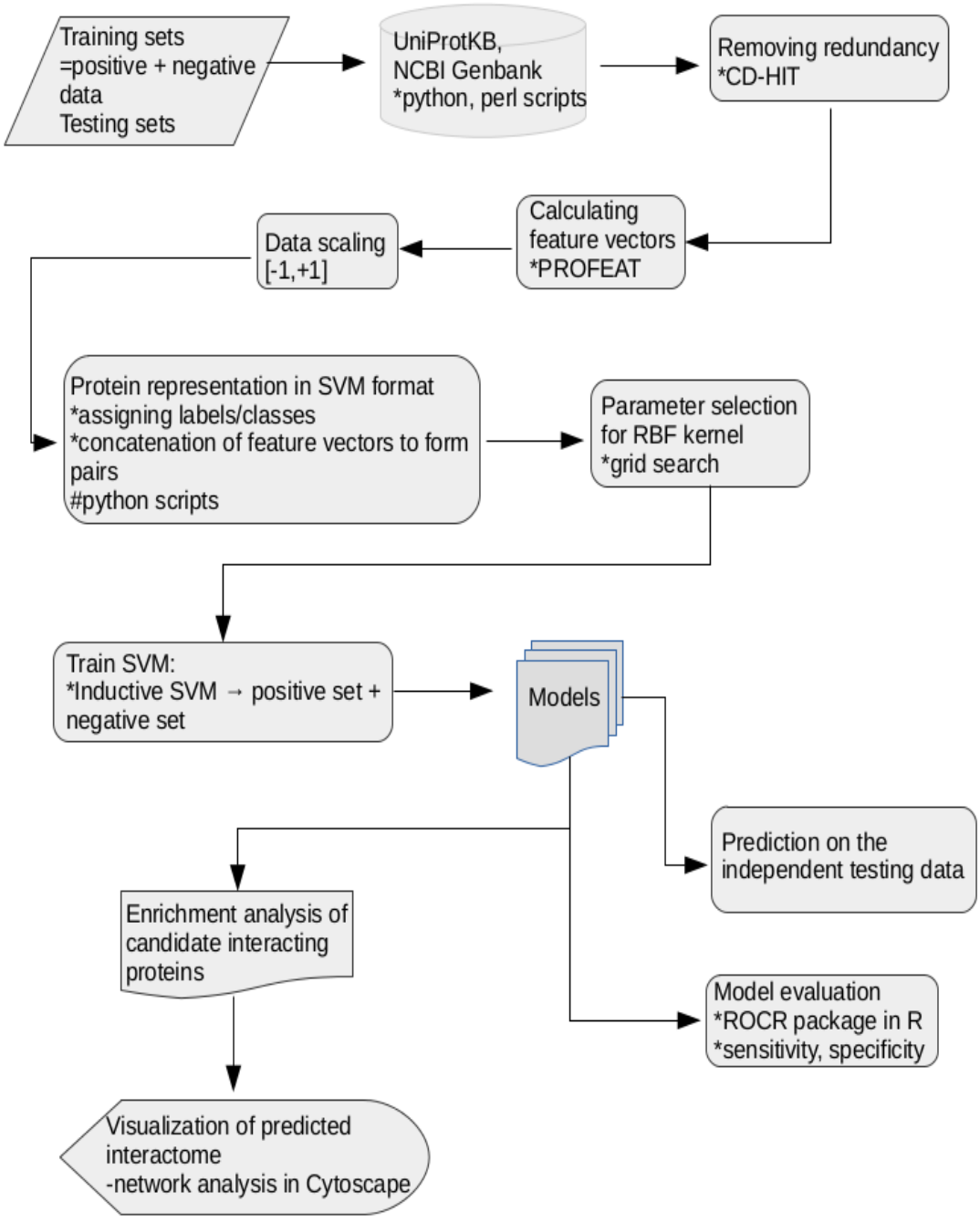

Figure 3.2 Schematic showing the summary of the methods.

.

3.5 Assessing the predicted protein interactions

3.5.1 Functional annotation of the predicted protein interactions between *T. parva* and bovine leukocyte

Bovine and parasite proteins predicted to interact were analyzed for significant enrichment of gene ontology terms using the GO::TermFinder (Boyle *et al*., 2004) and DAVID Bioinformatics Resources (http://david.abcc.ncifcrf.gov/). Bovine proteins were filtered by their location of expression, where they are likely to interact with secreted parasite proteins. Annotations depicting expression on cell surface, membrane proteins, cytoplasm and immune system involvement were considered. Bovine proteins with the following function and localization annotations were removed from the interactions data: 'mitochondria', 'ribosome', 'intracellular membrane-bounded organelle', 'nuclear chromatin', 'intracellular organelle', 'endoplasmic reticulum lumen', 'helicase activity', 'nuclease activity', 'nuclear body' and many others.

3.5.2 Gene expression and essentiality

In addition to filtering interaction data based on functional annotation, the predicted interaction was assessed by comparing to gene expression and essentiality. The transcriptome of schizont, the life stage of *Theileria parva* known to be center stage to east coast fever pathogenesis, has been fully sequenced and was used here. Differentially expressed *Theileria annulata* genes were searched for their orthologs

in *Theileria parva.* The two are closely related species, and both infect bovine leukocytes to induce changes that promote cell division and generate a neoplastic phenotype. *Theileria parva* genes essential for *in vivo* infection were also searched in literature. The predicted interactions were compared to genome-scale expression data sets describing gene expression changes that occur in *T. annulata* sporozoites infection (McKeever *et al*., 1997; Jensen *et al*., 2006; Jensen *et al*., 2008; Yamada *et al*., 2009). Data on bovine genes differentially expressed in T cells or macrophages during infection with *T. parva* or *T. annulata* was retrieved from NCBI's Gene Expression Omnibus database (GEO). Expression data was extracted from NCBI GEO Series Accession No. GSE4441. This was, however, challenging since available gene expression datasets measure expression in either the host or parasite but not both simultaneously.

3.6 Visualizing the predicted *T. parva* – bovine protein interactions

The predicted interaction was visualized and represented as an undirected graph of nodes and edges where nodes represented proteins and edges symbolized interactions between nodes. Network data files were prepared as *Simple Interaction Format* (.sif) and imported in to Cytoscape v. 2.8 software. For each *T. parva* protein and its predicted bovine partners, a value feature for degree (direct interactions) and for centrality were calculated in Cytoscape and distribution (scatter) plots for these two measures were made.

CHAPTER FOUR

RESULTS

4.1 Predicted protein-protein interactions

Twenty-three *T. parva* proteins out of the initial 211 were predicted to interact with 4001 bovine proteins and are outlined in Table 4.1. Two parasite proteins (Q4N068, Q4N069) were identified as cysteine proteinases, belonging to peptidase C1 family, and one protein (Q4N5N0) was identified as falcilysin with proteolytic activity and belonging to peptidase M16 family. A signal peptidase (Q4MYN7) was also identified, which is located on the parasite membrane with a transmembrane helix and has the ability to cleave hydrophobic, N-terminal signal or leader sequences from secreted and periplasmic proteins. Another parasite protein (Q9BH70) predicted to interact with bovine proteins was identified as Glutaredoxin-like protein.

Other *T. parva* proteins (13 in total) were identified as uncharacterized proteins or hypothetical proteins. Similarly, bovine proteins in the interaction data derived were functionally annotated and clustered. Enriched GO terms shared among the bovine proteins predicted to interact with *T. parva* proteins are listed in Table 1. Bovine proteins predicted to interact with *Theileria parva* parasites were also analyzed in the context of biological pathways (Table 2). Enriched signaling pathways included PI-3K, JAK/STAT and p38 MAPK. A large number of bovine proteins (about 541 proteins) predicted to interact with *T. parva* were associated with metabolic processes. *T. parva* protein Q4N3V0 was predicted to interact with bovine profilin

proteins (PFN2, PFN3), VIL1 and Destrin (actin depolymerizing factor). Bovine proteins located in- or those that function in -membrane bound organelles (including translation, RNA processing, protein degradation, ubiquitination metabolism and nuclear/cytoplasmic transport) were excluded from the interaction data.

Two *T. parva* proteins (Q4N8D0 and Q4N865) were predicted to interact with the bovine's AKN3 and FAM82B proteins located in the cytoskeleton. Another interaction was predicted between *T. parva* protein (Q4N3F7) and a bovine Rho GTPase protein (PLEKHG5), which has signal transducing activity including positive regulation of I-kappaB kinase in NF-kappaB signaling. Polymorphic immune-dominant molecule (PIM, Q27033), a membrane protein closely related to TaSP in *Theileria annulata*, was predicted to interact with tubulin proteins of cytoskeleton. *T. parva* protein TP03_0024 was predicted to interact with F-actin capping proteins including CAPZB, CAPZA1 and CAPZA2, while two other *T. parva* proteins (Q4N621 and Q4N841) were predicted to interact with bovine Toll like family of receptors (TLRs - TLR2, TLR4, TLR6 and TLR10) and IRAK1 protein complex known to stimulate the activity of NF-kB. Three bovine tubulin proteins - TUBB, TUBB2A, ATP5A1 - annotated to function as structural constituents of the cytoskeleton and microtubule were identified as putative targets of Q4N8J0, Q4N621, Q4N6T4, Q4N3M3, Q4MYG8 and Q4N8A9 *T. parva* proteins.

Table 4.1: *Theileria parva* proteins in the predicted host–parasite protein–protein interactions.

Parasite protein ID	Description
Q4N8A9, Q4N0T3, Q4N841, Q4N621, Q4N8J0, Q4N6T4, Q4N6C8, Q4N4C3, Q4N3M3, Q4N2S1, Q4MZ58, Q4MYF6, Q4N7V1, Q4N2R7, Q4MYG8, Q4MYM9, Q4N527	Uncharacterized protein; 'putative';
Q4N068, Q4N069	Cysteine proteases (peptidase C1)
Q4N5N0	Falcilysin (peptidase M16)
Q4MYN7	Signal peptidase
Q9BH70	Glutaredoxin-like protein*
P24724	HSP90

Table 4.2: Functions enriched in bovine proteins predicted to interact with *T. parva* proteins.

I) Molecular function of all bovine proteins predicted to interact with *T. parva*	
GO:0005488	Binding (calcium ion, protein, nucleic acid)
GO:0003824	Catalytic activity (hydrolase, deaminase, helicase, isomerase, ligase, lyase, oxidoreductase, transferase)
GO:0005215	Transporter activity (transmembrane transporter - amino acid, carbohydrate, lipid)
GO:0045182	Translation regulator activity (translation initiator, translation elongation)
GO:0005198	Structural molecule activity (structural constituent of cytoskeleton, structural constituent of ribosome)

GO:0004872	Receptor activity (G-PCR, ligand activated sequence specific DNA binding RNA pol II transcription factor activity, cytokine receptor, transmembrane receptor protein kinase, TNF activated receptor binding)
GO:0030234	Enzyme regulator activity (enzyme activator, enzyme inhibitor, kinase regulator, small GTPase regulator, phosphatase regulator, guanyl-nucleotide exchange factor activity)
GO:0001071	Nucleic acid binding transcription factor activity
GO:0000988	Protein binding transcription factor activity
GO:0016209	Antioxidant activity
II) Biological process of all bovine proteins predicted to interact with *T. parva*	
GO:0008152	Metabolic processes (primary metabolism, catabolic processes, biosynthetic processes)
GO:0006915	Apoptotic process (induction, negative regulation)
GO:0022610	Cell adhesion
GO:0065007	Biological regulation (homeostatic process)
GO:0071840	Cellular component organization or biogenesis
GO:0009987	Cellular process (cell communication, cell cycle, proliferation, chromosome segregation, cytokinesis)
GO:0032502	Developmental process (anatomical structure morphogenesis, cell differentiation, death)
GO:0008152	Immune system process (antigen processing and presentation, immune response, macrophage activation)
GO:0051179	Localization (RNA, transport)
GO:0032501	Multicellular organismal process
GO:0050896	Response to stimulus (endogenous and external stimulus, stress)
GO:0006968	Cell defense response

III) Cellular component of bovine proteins predicted to interact with *T. parva*	
GO:0030054	Cell junction
GO:0044464	Cell part (cytoplasm)
GO:0031012	Extracellular matrix
GO:0005576	Extracellular region
GO:0032991	Macromolecular complex (protein complex, ribonucleoprotein complex)
GO:0016020	Membrane (integral to membrane, mitochondrial inner membrane)
GO:0043226	Organelle (chromosome, cytoskeleton, mitochondrion, nucleus)

Table 4.3: Selected biological pathways and number of bovine protein in the corresponding pathway as identified using PANTHER Gene List analysis tool (Mi *et al*., 2013).

Pathway	Number of bovine proteins involved
Apoptosis signaling pathway (P00006)	41
Angiogenesis (P00005)	51
Ubiquitin proteasome pathway (P00060)	20
p53 pathway (P00059)	19
Wnt signaling pathway (P00057)	79
VEGF signaling pathway (P00056)	27

Pathway	Number of bovine proteins involved
Toll receptor signaling pathway (P00054)	23
TGF-beta signaling pathway (P00052)	16
PI3 kinase pathway (P00048)	17
Oxidative stress response (P00046)	14
PDGF signaling pathway (P00047)	41
Integrin signalling pathway (P00034)	45
Inflammation mediated by chemokine and cytokine signaling pathway (P00031)	69
Ras Pathway (P04393)	35
EGF receptor signaling pathway (P00018)	54
FGF signaling pathway (P00021)	46
Glycolysis (P00024)	15
Cytoskeletal regulation by Rho GTPase (P00016)	39
Cadherin signaling pathway (P00012)	46
p38 MAPK pathway (P05918)	21

4.2 Support vector machine models

A linear SVM model trained with datasets that were not scaled had more error rates and more learning iterations compared to a linear model trained with scaled features. Normalized feature vectors performed better than their non-normalized counterparts. Prediction results of classifying independent test set of known class with the inductive models based on linear and RBF kernels depict performance of the models (Table 4.4).

Table 4.4: Prediction results of classifying the independent test set with inductive SVM models.

	Accuracy	Recall	Precision	Sensitivity	Specificity	MCC
Linear kernel	77.83%	67.82%	80%	81.3%	78%	0.77
RBF kernel	86.22%	81.2%	84.72%	89.88%	84.39%	0.51

Using default learning parameters, the linear kernel-based model correctly predicted 467 out of 600 positive interactions and 379/600 negative interactions. The Linear kernel had a lower precision value, and thus a lesser capability of identifying correct positive samples. The RBF kernel-based SVM model performed better with improved accuracy, sensitivity and specificity. An optimal area under the curve (AUC) of 0.789 for the receiver operating characteristic (AUC-ROC) was obtained for the RBF kernel-based SVM model (Figure 4.1). AUC-ROC provides a single

measure of overall threshold independent accuracy by comparing specificity and sensitivity.

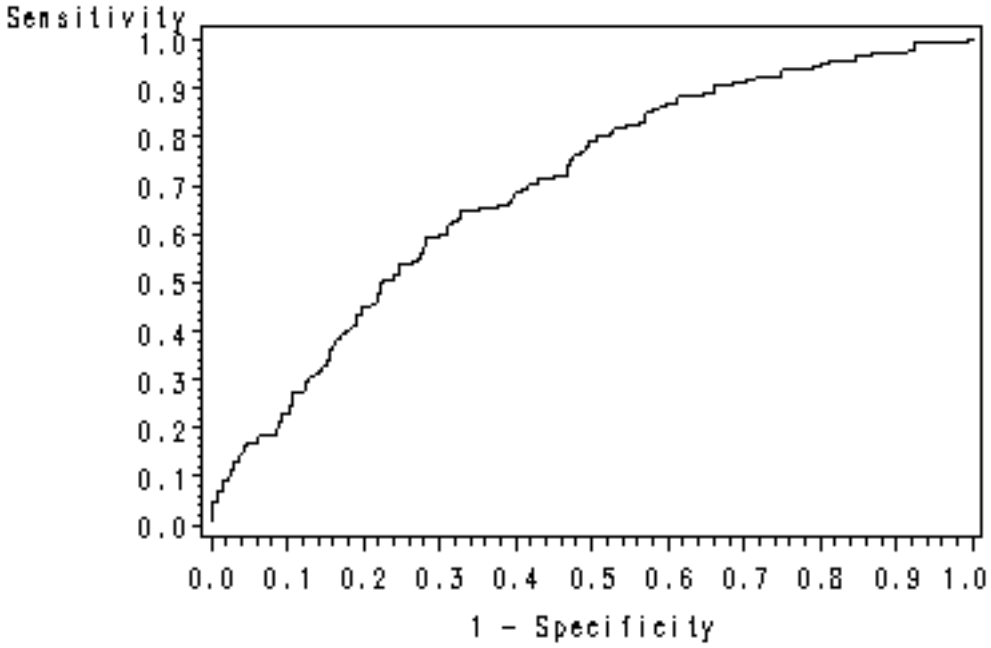

Figure 4.1: Performance of RBF kernel-based SVM model represented in a ROC curve (sensitivity vs. 1-specificity), AUC-ROC=0.79.

4.3 Comparison of predicted interactions to available gene expression and data

Theileria parasite invasion orchestrates a major reorganization of leukocyte gene expression networks. MMP9 was predicted to interact with Q4N841 parasite protein and FYN was predicted to interact with three parasite proteins, Q4N802, Q4N841 and Q4N0T3. NEMO, a central component of NF-kB signaling pathway, was predicted to interact with two *T. parva* proteins, Q4N0T3 and Q4N841. Bovine transcriptional regulators, JUN, JUNB, JUND, FOS and cMYC, were predicted to

interact with *T. parva*'s Q4N841 protein. Pro-cancer protein (PAK1) was predicted to interact with a parasite protein, and expression of the same protein was shown to be up-regulated in *Theileria* infection (Durrani *et al*., 2012). IKK-beta catalytic subunit, essential for NF-kB activation in response to pro-inflammatory stimuli, was predicted to interact with Q4N0T3 and Q4N841 *T. parva* proteins. Parasite proteins TP02_0244 and PIM were previously identified as the most abundant schizont proteins with corresponding high level of transcription (Bishop *et al*., 2005). These proteins were predicted to interact with numerous bovine proteins.

4.4 Network analysis and visualization

Predicted interactions were presented as undirected network or graph (Figure 4.2), which was analyzed for topological degree and between-ness centrality distributions using Cytoscape's network analysis algorithm. A large fraction of proteins in the predicted interaction network participated in fewer number of interactions, i.e. small degree and a small fraction of proteins were seen as high degree nodes or hubs. 1822 bovine proteins had at least two or more interacting parasite proteins. Seven different parasite proteins were predicted to interact with the same host protein and 17 bovine actin proteins were found to have most interactions with different parasite proteins. *T. parva* HSP90-like protein (P24724) was predicted to interact with a large number of bovine proteins. In this context, HSP90 proteins may function as hubs (those involved in many interactions) while actin proteins may function as bottlenecks (those central to many pathways).

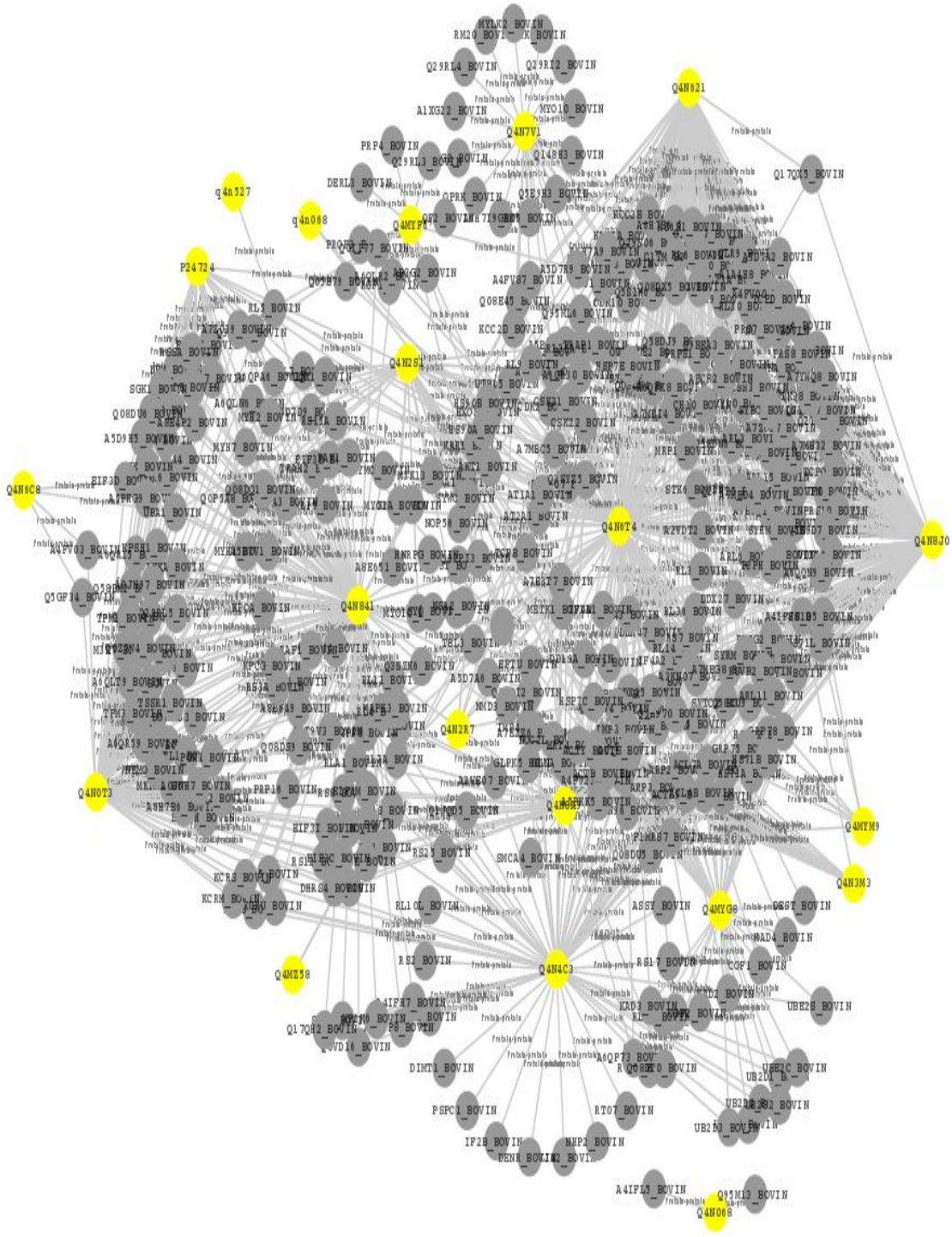
Q4N621
Q4N7V1
q4n527
q4n088
P24724
Q4N2S4
Q4N6C8
Q4N0T4
Q4N8J0
Q4N841
Q4N2R7
Q4N0T3
Q4MYM9
Q4N3M3
Q4MYG8
Q4N4C3
Q4MZ58
Q4N068

Figure 4.2 Predicted interactions presented as an undirected network or graph. *T. parva* proteins are shaded in yellow, while bovine proteins are shaded in gray. Edges or lines are seen connecting different nodes.

No self-interactions or self loops were reported. Network clustering coefficient, the average of the clustering coefficients for all nodes in the network, was estimated at 0.101, while network diameter and average shortest path length were estimated at 11 and 12.698570 (98%) respectively. Path length is the number of edges forming a path and multiple paths may connect two given nodes. Network diameter is the largest distance (shortest path length) between two nodes. Average number of neighbors and network heterogeneity were estimated at 6.445 and 4.819 respectively. The two parameters relate to neighborhood, where the former indicates the average connectivity of a node in the network and the latter reflects the tendency of a network to contain hub nodes. A correlation between node degree and between-ness for each predicted pathogen-host interaction is shown in Figure 4.3.

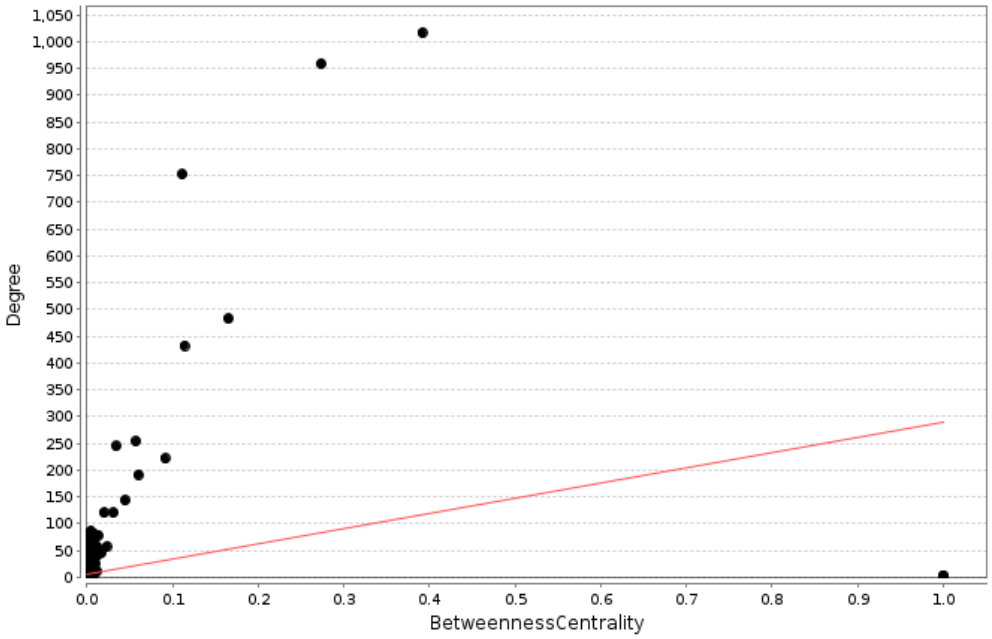

Figure 4.3 Degree – between-ness centrality correlation graph of the predicted interaction network. Each point in the graph represents a predicted interacting protein pair in the PIN.

CHAPTER FIVE

DISCUSSION, CONCLUSION, RECOMMENDATION

5.1 Discussion

This study presents the use of protein sequence composition, physico-chemical properties, and homolog-based information to predict protein interactions *in silico*. Interactions were also inferred from related pathogen-host systems, since orthologous genes are expected to retain similar function after speciation. Bovine proteins predicted to interact with *T. parva* were annotated in cellular component terms that are consistent with known mechanisms of infections, including cytoplasm, plasma membrane, extrinsic to the cell, cytoskeleton and cell junction. These mechanisms and biological pathways are illustrated in Figures 2.1 and 2.2 for reader's visualization and reference. Regulation of cell proliferation, apoptosis and tumor invasion is parasite-dependent changes in the bovine host cell described extensively by Heussler *et al*. (2002); Heussler *et al.* (2001a); Lizundia *et al*. (2006), and these biological processes were associated with the proteins predicted to interact with *T. parva.* These studies are useful for reference on the specific details of the biological processes mentioned above. Other predicted interactions were related to translation and gene regulation, immune system process, cell development and differentiation, metabolic processes and apoptosis, consistent with transformation phenotype (Shiels *et al*., 2006; Dobbeleare & Heussler, 1999; Dobbeleare & Kuenzi, 2004). 'Metabolism' was the most GO term shared among bovine proteins

highlighting the possible association of *T. parva* with bovine metabolic processes suggesting parasite's adaptation mechanism. Despite this being a general term, *T. parva* has been shown to have a reduced mitochondrial genome with absent biosynthetic pathways (Gardner *et al*., 2005) implying that the parasite may be dependent on the host cell for its metabolic needs. A large number of bovine proteins were also associated to cellular processes including cytokinesis, which is linked to parasite cell division through tight association with the leukocyte mitotic apparatus (Sager *et al*., 1997). NEMO was previously reported as transcriptionally silent in non-infected cells and may represent genes whose expression is important to the *Theileria* infected cell (Machado *et al*., 2000).

T. parva schizont lies free in the host cell cytoplasm where proteins containing signal peptide/anchor or trans-membrane domains are of critical interest since they may be secreted into the host cell or anchored into the plasma membrane and have the potential to interact with host cell components. Akt/PKB anti-apoptotic proteins predicted to interact to interact with *T. parva* are down-regulated on drug-mediated parasite elimination, a process that coincides with growth arrest of host cell, increased caspase activity followed by programmed cell death (Guergnon *et al*., 2003; Heussler *et al*., 2001b). In addition, PI-3K stimulation results in the activation of serine/threonine kinase Akt which regulates glucose metabolism (Dobbeleare & Kuenzi, 2004; Baumgartner *et al*., 2000), that may be essential for parasite survival and proliferation.

Glutaredoxin-like protein is involved in cell redox homeostasis, and this protein homolog had earlier been described as a nuclear encoded 17-kDa secretory or membrane protein containing a signal sequence for endoplasmic reticulum membrane association (Ebel *et al*., 1997). HSP90 parasite protein is a molecular chaperone, heat shock protein (Hsp90) that promotes proper maturation (protein folding), structural maintenance and regulation of proteins involved in cell cycle control and signal transduction. Bovine prolifin proteins are associated with actin organization in the cytoskeleton, cell migration and regulation of apoptosis. Capping proteins are known to regulate cell morphology and cytoskeletal organization (Schneider *et al*., 2007). Previously reported bovine genes whose expression is altered at RNA or protein level during *T. parva* infection (Kinnaird *et al*., 2013) were predicted to interact with the schizont derived proteins. These genes are linked to metastasis and proliferation of infected cells and including Matrix Metalloproteinase 9 (MMP9) (Baylis *et al*., 1995; Adamson *et al*., 2000) and FYN proteins (Src kinase, Baumgartner, 2011).

Theileria's TaSP proteins have been shown to co-localize and physically interact with tubulin proteins to regulate microtubule assembly (Seitzer *et al*., 2010). Their association with host microtubule network could enable the parasite to attach to the mitotic spindle apparatus and possibly play a role in parasite distribution into daughter host cells (von-Schubert *et al.*, 2010). *T. parva*'s interaction with host's Epidermal Growth Factor (EGF) signaling pathway could play a critical role in the regulation of cell growth, proliferation, and differentiation. Toll Like Receptors

(TLRs) elicit conserved pathways that culminate in the activation of NF-kB and AP1 transcription factors (Kawai & Akira, 2006). Interaction between *T. parva* and host cell trans-membrane TLRs may indicate pathogen's obstruction of inflammatory response and pathogen clearance. The interaction may counteract TLR detection by interfering with TLR signaling, keeping immune cells in an inactive state and rendering them refractory to subsequent TLR stimulation.

The *Theileria* schizont induces constitutive NF-κB activation by recruiting the IKK complex to its surface. At the schizont surface, IKK is activated and induces the phosphorylation and subsequent proteasomal degradation of IκBα, the cytoplasmic inhibitor of NF-κB. In the nucleus, NF-κB, together with other transcription factors, induces the expression of anti-apoptotic proteins (Dobbelaere & Kuenzi, 2004). In this sense, predicted interactions between *T. parva* proteins and IKK may contribute to protection against apoptosis at different levels. Transcriptional regulators were previously shown as highly up regulated in *Theileria* infection. FOS and JUN family members heterodimerise to form the AP1 complex that is constitutively activated in *Theileria* infected leukocytes (Chaussepied *et al*., 1998).

Pathogens interact with host proteins that are hubs (those involved in many interactions) or bottlenecks (those central to many pathways) that control cell critical processes (Yu *et al*., 2007). In the predicted PPI, proteins in the actin and HSP90 families were observed as targets for different parasite effectors and thus identified as high degree nodes. Actin proteins are involved in various types of cell motility and could be manipulated to promote infiltration and invasion of infected cells. Hsp90

are molecular chaperones that influence phosphorylation and activity of a wide range of key signaling proteins. Tubulin-beta proteins were also predicted to interact with multiple parasite proteins. These proteins polymerize into microtubules enabling the cell to undergo cell division.

Linear kernel had a lower precision value, and thus a lesser capability of identifying correct positive samples. A specificity of ~85% and a sensitivity of ~90% for the RBF-based kernel classifier imply that the classifier identified ~85% of the non-interacting test protein pairs and 90% interacting test protein pairs, correspondingly. Considering the inherent error rates from training and testing data, RBF kernel-based model was chosen as better performing, more suitable and more accurate. The performance of RBF kernel-based SVM model improved with increased accuracy, sensitivity and specificity.

Combining different evaluation metrics tends to provide comprehensive assessment of classification on imbalanced datasets. An optimal area under the curve (AUC) value of 0.789 for the receiver-operating characteristic (AUC-ROC) was reported for the RBF kernel-based SVM model. AUC-ROC is an important index that provides a single measure of overall threshold independent accuracy. Prediction accuracies achieved show good performance of the SVM classifiers on unseen data. However, the classifiers identified non-interacting (negative) proteins poorly compared to interacting proteins according to the specificity scores, likely due to the nature of the negative test data.

5.2 Conclusion

At molecular level, interactions between a pathogen and its host play critical roles in initiating infection and a successful pathogenesis. The goal of this study was to predict interactions between bovine and *T. parva* proteins that are likely to be critical in east coast fever pathogenesis. The predicted interactions are associated with molecular functions and host pathways previously reported as targeted or triggered upon infection by the parasite. Molecular functions and pathways in the predicted interactions provide insights into the possible host pathways targeted by the parasite, interactions that could be essential for survival of the transformed cell and successful dissemination of the parasite. This study has provided a general overview of the landscape of bovine proteins interacting with the schizont stage of *T. parva*. By correctly identifying known, experimentally validated protein interactions (90% sensitivity), this study validates the use of compositional features, physico-chemical properties of protein sequences and application of predictive models for inferring inter-molecular interactions. These use of these features support future direction for use of computational methods augmented with additional transferred knowledge (homolog information). Thus computational methods can be used to predict or transfer protein interactions.

An undirected PPI network between host and parasite proteins was constructed to provide further insights into manipulation of host cellular programs. The study identified likely hubs (those involved in many interactions) and bottlenecks (those central to many pathways) in the predicted interaction network, which included host

actin proteins and parasite's heat shock protein. Parasite proteins were discovered to interact with hubs and bottlenecks in the constructed network.

The study also validated use of homolog-based information (knowledge transfer from related pathogen systems) to infer protein interactions in host-pathogen systems. Different filtering techniques were considered for assessing the feasibility of the interactions under an in vivo condition and consequently decreasing the false positives, including biologically feasible interactions including integration of expression and sub-cellular localization data, gene ontology annotations, and presence or absence of translocational signals in parasite's proteins. This study highlights potential interactions, which is useful in limiting future experimental scope in *Theileria* pathogenesis. This study was limited by scarcity of information on proteins specific to *T. parva* infected bovine leukocytes. Nonetheless, predictions inferred here should complement future experimental efforts for elucidating bovine-*T. parva* interactions. Putative PPIs predicted in this study should further be validated by time-series microarray and yeast hybrids experiments.

5.3 Recommendations

The main challenge was that truly interacting host-pathogen PPI data and the relevant feature information is scarce or available for few pathogen-host systems. Equally, lack of verified non-interacting protein pairs and missing feature information for proteins was a challenge. The study recommends application of interspecies transcriptomics data for better construction of bovine-*T. parva* PPI network. Putative PPIs predicted in this study should further be validated by time-

series microarray and yeast hybrids experiments; and it would be useful to design better methods for inference of non-interacting proteins.

REFERENCES

Adamson, R., Logan, M., Kinnaird, J., Langsley, G., and Hall, R. (2000). Loss of matrix metalloproteinase 9 activity in *Theileria annulata*-attenuated cells is at the transcriptional level and is associated with differentially expressed AP-1 species. *Molecular Biochemistry Parasitology, 106*, 51-61.

Akoolo, L., Pellé, R., Saya, R., Awino, E., Nyanjui, J., Taracha, E. L., Kanyari, P. ... and Graham, S. P. (2008). Evaluation of the recognition of *Theileria parva* vaccine candidate antigens by cytotoxic T lymphocytes from Zebu cattle. *Veterinary Immunology Immunopathology, 15*, 216-221.

Barh, D., Gupta, K., Jain, N., Khatri, G., León-Sicairos, N., Canizalez-Roman, A., Smith, J.C. ... and Harmit, R. (2013). Conserved host-pathogen PPIs. Globally conserved inter-species bacterial PPIs based conserved host-pathogen interactome derived novel target in *C. pseudotuberculosis, C. diphtheriae, M. tuberculosis, C. ulcerans, Y. pestis*, and *E. coli* targeted by Piper betel compounds. *Integrative Biology*, *5*, 495–509.

Baumgartner, M., Chaussepied, M., Moreau, M. F., Werling, D., Davis, W. C., Garcia, A. ... and Cenilez-Rowat, P. (2000). Constitutive PI3-K

activity is essential for proliferation, but not survival, of *Theileria parva*-transformed B cells. *Cell Microbiology*, *2*, 329-339.

Baumgartner, M. (2011). *Theileria annulata* promotes Src kinase-dependent host cell polarization by manipulating actin dynamics in podosomes and lamellipodia. *Cell Microbiology, 13*, 538-553.

Baylis, H. A., Megson, A., and Hall, R. (1995). Infection with *Theileria annulata* induces expression of matrix metalloproteinase 9 and transcription factor AP-1 in bovine leucocytes. *Molecular Biochemistry Parasitology, 69*, 211-222.

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

Ben-Hur, A., Ong, C. S., Sonnenburg, S., Scholkopf, B. and Ratsch, G. (2008). Support vector machines and kernels for computational biology. *PloS Computational Biology, 4(10)*, e1000173.

Bishop, R., Geysen, D., Spooner, P., Skilton, R., Nene, V., Dolan, T., and Morzaria, S. (2001). Molecular and immunological characterization of *Theileria parva* stocks which are components of the 'Muguga cocktail' used for vaccination against East Coast fever in cattle. *Veterinary Parasitology*, *94*, 227–237.

Bishop, R., Geysen, P. D., Skilton, R., Odongo, D., Nene, V., Sang, P. R., ... and Farady, J. (2002). Genomic polymorphism sexual recombination and molecular epidemiology of *Theileria parva*. *Molecular Biochemistry Parasitology, 9,* 78-96.

Bishop, R., Musoke, A., Mozaria, S., Gardner, M. and Nene, V. (2004). *Theileria*: intracellular protozoan parasites of wild and domestic ruminants transmitted by ixodid ticks. *Parasitology*, *129*, S271-S283.

Bishop, R., Shah, T., Pelle, R., Hoyle, D., Pearson, T., Haines, L., Brass, A., ... and Michaels, J. (2005). Analysis of the transcriptome of the protozoan *Theileria parva* using MPSS reveals that the majority of genes are transcriptionally active in the schizont stage. *Nucleic Acids Research*, *33(17)*, 5503–5511.

Boyle, E.I., Weng, S., Gollub, J., Jin, H., Botstein, D., Cherry, J.M., and Sherlock, G. (2004). GO::TermFinder - open source software for accessing gene ontology information and finding significantly enriched gene ontology terms associated with a list of genes. *Bioinformatics, 20*, 3710-3715.

Bradford, J. R. and Westhead, D. R. (2005). Improved prediction of protein–protein binding sites using a support vector machines approach. *Bioinformatics, 21*, 1487–1494.

Brown, C. G., Stagg, D. A., Purnell, R. E., Kanhai, G. K., and Payne, R. C. (1973). Infection and Transformation of Bovine Lymphoid Cells in vitro by Infective Particles of *Theileria parva*. *Nature*, *245*, 101-103.

Burger, L. and Nimwegen, E. (2008). Accurate prediction of protein-protein interactions from sequence alignments using a Bayesian method. *Molecular Systems Biology*, *4*, 165.

Chaussepied, M., Lallemand, D., Moreau, M. F., Adamson, R., Hall, R., and Langsley, G. (1998). Upregulation of Jun and Fos family members and permanent JNK activity lead to constitutive AP-1 activation in *Theileria*-transformed leukocytes. *Molecular Biochemistry Parasitology, 94*, 215-226.

Cristianini, N. and Shawe-Taylor, J. (2000). Support Vector Machines and other Kernel-based learning methods. *IEEE, 5*, 128-133.

Dessauge, F., Lizundia, R., Baumgartner, M., Chaussepied, M. and Langsley, G. (2005a). Taking the Myc is bad for *Theileria*. *Trends in Parasitology, 21*, 377–385.

Dobbelaere, D. and Heussler, V. (1999). Transformation of leukocytes by *Theileria parva* and *Theileria annulata. Annual Review Microbiology, 53,* 1–42.

Dobbelaere, D., Fernandez, P., Machado, J., Botteron, C., Heussler, V. (1999). Interference by the intracellular parasite *Theileria parva* with T-cell signal transduction pathways induces transformation and protection against apoptosis. *Veterinary Immunology Immunopathology, 72,* 95-100.

Dobbelaere, D. A. and Kuenzi, P. (2004). The strategies of the Theileria parasite: a new twist in host-pathogen interactions. *Current Opinion in Immunology, 16,* 524–530.

Durrani, Z., Weir, W., Pillai, S., Kinnaird, J., and Shiels, B. (2012). Modulation of activation-associated host cell gene expression by the apicomplexan parasite *Theileria annulata*. *Cell Microbiology, 14*, 1434–1454.

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

Ebel, T., Middleton, J. F. S., Frisch, A. and Lipp, J. (1997). Characterization of secretory type *Theileria parva* glutaredoxin homologue identified by novel screening procedure. *Journal of Biological Chemistry, 272*, 3042-3048.

Gachohi, J., Skilton, R., Hansen, F., Ngumi, P., and Kitala, P. (2012). Epidemiology of East Coast Fever (*Theileria parva* infection) in Kenya: past, present and the future. *Parasites and Vectors, 5*, 194.

Garcia-Garcia, J., Guney, E., Aragues, R., Planas-Iglesias, J., and Oliva, B. (2010). BiANA: a software framework for compiling biological interactions and analyzing networks. *BMC Bioinformatics, 11*, 56.

Gardner, M. J., Bishop, R., Shah, T., de Villiers, E. P., Carlton, J. M., Sanghi-Majh, A. R., ... and Mannkly, J. T. (2005). Genome sequence of *Theileria parva*, a bovine pathogen that transforms lymphocytes. *Science 309*, 134–137.

Geysen, D., Bishop, R., Skilton, R., Dolan, T. T. and Morzaria, S. (1999). Molecular epidemiology of *Theileria parva* in the field. *Tropical Medicine and International Health, 4*, A21– A27.

Gonzalez, M. W. and Kann, M. G. (2012). Protein Interactions and Disease. *PLoS Computional Biology, 8(12)*, e1002819.

Graham S. P., Pelle, R., Honda,Y., Mwangi, D. M., Tonukari, N. J., Yamage, M., ... and and Brest, M. (2006). *Theileria parva* candidate vaccine antigens recognized by immune bovine cytotoxic T lymphocytes. *Proceedings of National Academy of Science*, *103*, 3286–3291.

Graham, S.P., Pellé, R., Yamage, M., Mwangi, D.M., Honda, Y., Mwakubambanya, R.S., ... and Phillips, J. T. (2008). Characterization of the fine specificity of bovine CD8 T-cell responses to defined antigens from the protozoan parasite *Theileria parva*. *Infection and Immunity, 76*, 685–694.

Guergnon, J., Dessauge, F., Langsley, G. and Garcia, A. (2003). Apoptosis of *Theileria*-infected lymphocytes induced upon parasite death involves activation of caspases 9 and 3. *Biochimie*, *85*, 771–776.

Heussler, V. T., Machado, J. Jr., Fernandez, P. C., Botteron, C., Chen, C.-G., Pearse, M. J. ... and Dobbelaere, D. A. E. (1999). The intracellular parasite *Theileria parva* protects infected T cells from apoptosis. *Proceedings of the National Academy of Sciences*, *96*, 7312–7317.

Heussler, V. T., Ku¨enzi, P. and Rottenberg, S. (2001a). Inhibition of apoptosis by intracellular protozoan parasites. *International Journal for Parasitology*, *31*, 1166–1176.

Heussler, V. T., Ku¨enzi, P., Fraga, F., Schwab, R. A., Hemmings, B. A. and Dobbelaere, D. A. E. (2001b). The Akt/PKB pathway is constitutively activated in *Theileria*-transformed leucocytes, but does not directly control constitutive NF-kB activation. *Cellular Microbiology, 3,* 537–550.

Heussler, V. T., Rottenberg, S., Schwab, R., Kuenzi, P., Fernandez, P.C., McKellar, S., ... and Dobbelaere, D.A.E. (2002). Hijacking of host cell IKK signalosomes by the transforming parasite *Theileria*. *Science, 298,* 1033–1036.

Heussler, V., Sturm, A. and Langsley, G. (2006). Regulation of host cell survival by intracellular *Plasmodium* and *Theileria* parasites. *Parasitology*, *132*, S49-S60.

Jensen, K., Talbot, R., Paxton, E., Waddington, D., and Glass, E. J. (2006). Development and validation of a bovine macrophage specific cDNA microarray. *BMC Genomics, 7,* 224.

Jensen, K., Paxton, E., Waddington, D., Talbot, R., Darghouth, M. A., Strum, K., ... and Lai, J. M. (2008). Differences in the transcriptional responses induced by *Theileria annulata* infection in bovine monocytes derived from resistant and susceptible cattle breeds. *International of Journal of Parasitology, 38,* 313–325.

Joachims, T. (1999). *Making large-Scale SVM Learning Practical*. Advances in Kernel Methods - Support Vector Learning. *IEEE, 8*, 123-130.

Joachims, T. (2002). *Learning to Classify Text Using Support Vector Machines*. *IEEE, 1*, 87-97.

Jothi, R., Cherukuri, P. F., Tasneem, A., and Przytycka, T. M. (2006). Co-evolutionary analysis of domains in interacting proteins reveals insights into domain-domain interactions mediating protein–protein interactions. *Journal of Molecular Biology*, *362(4),* 861–875.

Kawai, T. and Akira, S. (2006). TLR signaling. *Cell Death and Differentiation, 13,* 816-825.

Kawashima, S. and Kanehisa, M. (2000). AAindex: amino acid index database. *Nucleic Acids Research, 28*, 374.

Kinnaird, J. H., Weir, W., Durrani, Z., Pillai, S. S., Baird, M. Mssai, P. H., ... and Lee, J. (2013). A Bovine Lymphosarcoma Cell Line Infected with *Theileria annulata* Exhibits an Irreversible Reconfiguration of Host Cell Gene Expression. *PLoS ONE 8(6),* e66833.

Ku¨enzi, P., Schneider, P. and Dobbelaere, D. A. (2003). *Theileria parva*-transformed T cells show enhanced resistance to Fas/Fas ligand-induced apoptosis. *Journal of Immunology, 171*, 1224–1231.

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

Li, W. and Godzik, A. (2006). CD-HIT: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics, 22*, 1658-9.

Lizundia, R., Sengmanivong, L., Guergnon, J., Mu¨ller, T., Schnelle, T., Langsley, G. and Shorte, S. L. (2005). Use of micro-rotation imaging to study JNK-mediated cell survival in *Theileria parva*-infected B-lymphocytes. *Parasitology 130*, 629–635.

Lizundia, R., Chaussepied, M., Huerre, M., Werling, D., Di Santo, J. P., Guergnon, J., ... and Langey, G. (2006). c-Jun NH2-terminal kinase/c-Jun signaling promotes survival and metastasis of B lymphocytes transformed by *Theileria*. *Cancer Research, 66*, 6105–6110.

Luder, C. G., Stanway, R. R., Chaussepied, M., Langsley, G., and Heussler, V. T. (2009). Intracellular survival of apicomplexan parasites and host cell modification. *International Journal of Parasitology, 39*, 163-173.

Machado, J. Jr, Fernandez, P. C., Baumann, I., and Dobbelaere, D. A. (2000). Characterisation of NF-kappa B complexes in *Theileria parva*-transformed T cells. *Microbes and Infections, 2*, 1311–1320.

MacHugh, N. D., Connelley, T., Graham, S. P., Pelle, R., Formisano, P., Taracha, E. L., ... and Morrison, W. I. (2009). CD8+ T-cell responses to *Theileria parva* are preferentially directed to a single dominant antigen: Implications for parasite strain-specific immunity. *European Journal of Immunology*, *39*, 2459–2469.

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

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

McKeever, D. J., Nyanjui, J. K., and Ballingall, K. T. (1997). In vitro infection with *Theileria parva* is associated with IL10 expression in all bovine lymphocyte lineages. *Parasite Immunology, 19*, 319-324.

Mi, H., Muruganujan, A., Casagrande, J. T. and Thomas, P. D. (2013). Large-scale gene function analysis with the PANTHER classification system. *Nature Protocols 8*, 1551- 1566.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., Ellis, S. A., ... and Burrells, A. (2002). Network motifs: Simple building blocks of complex networks. *Science, 298*, 824–827.

Pelle, R., Graham, S. P., Njahira, M. N., Osaso, J., Saya, R. M., Millis, S. T., ... and Nene, V. (2011). Two *Theileria parva* CD8 T cell antigen genes are more variable in buffalo than cattle parasites, but differ in pattern of sequence diversity. *Plos ONE*, 6(4), e19015.

Plattner, F. and Soldati-Favre, D. (2008). Hijacking of host cellular functions by the Apicomplexa. *Annual Review Microbiology, 62*, 471-87.

Qi, Y., Bar-Joseph, Z. and Klein-Seetharaman, J. (2006). Evaluation of different biological data and computational classification methods for use in protein interaction prediction. *Protein: Structure, Function, and Bioinformatics, 63,* 490–500.

Rao, H. B., Zhu, F., Yang, G. B., Li, Z. R. and Chen, Y. Z. (2011). Update of PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. *Nucleic Acids Research, 39,* W385-90.

Sager, H., Davis, W. C., Dobbelaere, D. A. and Jungi, T. W. (1997). Macrophage parasite relationship in *theileriosis*. Reversible phenotypic and functional de-differentiation of macrophages infected with *Theileria annulata*. *Journal of Leukocyte Biology*, *6*, 459–468.

Schaumburg, F., Hippe, D., Vutova, P. and Luder, C. G. K. (2006). Pro- and anti-apoptotic activities of protozoan parasites. *Parasitology*, *132*, S69-S85.

Schmuckli-Maurer, J., Kinnaird, J., Pillai, S., Hermann, P., McKellar, S., Kuman-Illis, S., ... and Koi, J. G. (2010). Modulation of NF-kappaB activation in *Theileria annulata*-infected cloned cell lines is associated with detection of parasite-dependent IKK signalosomes and disruption of the actin cytoskeleton. *Cell Microbiology*, *12*, 158–173.

Schneider, I., Haller, D., Kullmann, B., Beyer, D., Ahmed, J. S., and Seitzer, U. (2007). Identification, molecular characterization and subcellular localization of a *Theileria annulata* parasite protein secreted into the host cell cytoplasm. *Parasitology Research*, *101*, 1471-1482.

Seitzer, U., Gerber, S., Beyer, D., Dobschanski, J., Kullmann, B., Haller, D., ... and Tyrrell, A. M. (2010). Schizonts of *Theileria annulata* interact with the microtubuli network of their host cell via the membrane protein TaSP. *Parasitology Research, 106*, 1085-1102.

Shaw, M. K. (1997). The same but different: the biology of *Theileria* sporozoite entry into bovine cells. *International Journal of Parasitology, 27*, 457 – 474.

Shaw, M. K. (2003). Cell invasion by *Theileria* sporozoites. *Trends in Parasitology, 19*, 2–6.

Shen, J., Zhang, J., Luo, X., Zhu, W., Yu, W., Chen, K., ... and Jiang, H. (2007). Predicting protein-protein interactions based only on sequences information. *PNAS Biophysics*, *104(11)*, 4337-4341.

Shiels, B., Langsley, G., Weir, W., Pain, A., McKellar, S. and Dobbelaere, D. (2006). Alteration of host cell phenotype by *Theileria annulata* and *Theileria parva*: mining for manipulators in the parasite genomes. *International Journal of Parasitology, 36*, 9–21.

Shoemaker, B. A. and Panchenko, A. R. (2007a). Deciphering protein–protein interactions. Part I. Experimental techniques and databases. *PLoS Computional Biology, 3(3)*, e42.

Shoemaker, B. A and Panchenko, A. R. (2007b). Deciphering protein–protein interactions. Part II. Computational methods to predict protein and domain interaction partners. *PLoS Computional Biology, 3(4),* e43.

Skrabanek, L., Saini, H. K., Bader, G. D. and Enright, A. J. (2008). Computational prediction of protein-protein interaction. *Molecular Biotechnology, 38,*1-17.

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

Tato, C. M. and Hunter, C. A. (2002). Host-pathogen interactions: subversion and utilization of the NF-kB pathway during infection. *Infection and Immunity, 70(7),* 3311-3317.

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

von-Schubert, C., Xue, G., Schmuckli-Maurer, J., Woods, K. L., Nigg, E. A., and Dobbelaere, D. A. (2010). The transforming parasite *Theileria* co-opts host cell mitotic and central spindles to persist in continuously dividing cells. *PLoS Biology, 8,* 43.

Yamanishi, Y., Vert, J. P. and Kanehisa, M. (2004). Protein network inference from multiple genomic data: a supervised approach. *Bioinformatics*, *20*, 363–370.

Yamada, S., Konnai, S., Imamura, S., Simuunza, M., Chembensofu, M., Chota, A., ... and Ohashi, K. (2009). Quantitative analysis of cytokine mRNA expression and protozoan DNA load in *Theileria-parva* infected cattle. *Journal of Veterinary Medical Sciences, 71(1),* 49–54.

Yu, H., Luscombe, N. M., Lu, H. X., Zhu, X., Xia, Y., Linnama, S., ... and Long, L. H. (2004). Annotation transfer between genomes: protein-protein interologs and protein–DNA regulogs. *Genome Research*, *14*, 1107–1118.

Yu, H., Kim, P. M., Sprecher, E., Trifonov, V. and Gerstein, M. (2007). The importance of bottlenecks in protein networks: Correlation with gene essentiality and expression dynamics. *PLoS Computational Biology*, *3*(4), e59.

Zhang, L. V., Wong, S. L., King, O. D., Roth, F. P. (2004). Predicting co-complexed protein pairs using genomic and proteomic data integration. *BMC Bioinformatics, 5,* 38.

APPENDICES

Appendix 1: Fetching Sequences from UniProtKB (script1.py)

```
#!/usr/bin/python

#retrieve fasta sequences for training, July 26 / 13

import sys
def retrieve_seq(line):
    from Bio import Entrez
    from Bio import SeqIO
    sys.stdout = file(sys.argv[2],"a")
    Entrez.email = "e.kamau@cgiar.org"
    handle = Entrez.efetch(db = "protein", rettype = "fasta", retmode = "text", id = line)
    for seq_record in SeqIO.parse(handle, "fasta"):
        print '>',seq_record.id
        print seq_record.seq
    handle.close()

seq_ids = file(sys.argv[1],"r")
for line in seq_ids:
    retrieve_seq(line)
```

Appendix 2: Making random protein sequences (script2.R)

```
# A script to generate random amino acid or nucleotide sequences from a vector with replacement.

# NB: We require the "seqinr" package which will be used in the final stages to write a fasta file.
# Uncomment the line below to install that package.
install.packages("seqinr");

# Load the package
library(seqinr);

# The 20 standard amino acids
aa <- c("A","R","N","D","C","Q","E","G","H","I","L","K","M","F","P","S","T","W","Y","V");

# If we wanted to generate random nucleotide sequences this is the vector we would use instead of the
"aa" above.
bases <- c("A", "T", "G", "C");

# Create fasta format style header names for each sequence
seqnames <- paste(">Random Amino Acid Seq no: ", 1:1500,sep="");

# Generate the sequences. The most important function here is the sample function.
# The Sample function: This function shuffles the contents of a vector into a random sequence while
maintaining all
# the numerical values intact. It is extremely useful for randomization in experimental design,
# in simulation and in computationally intensive hypothesis testing. In this instance, I have used it
with the
# replacement option set to true, meaning some residues will be repeated in the random sequence.
seqs <- sapply(1:1500, function(x) paste(sample(aa, size=600, replace=T), collapse=""));

# Write the sequences to file: This is the only place we use one of the functions of the
# "seqinr" package - its "writeLines" function that writes sequences in fasta format.
# NB: The "aa_seqs.fasta" file is writen in the current directory as can be determined by typing "getwd
()"
writeLines(as.vector(t(cbind(seqnames, seqs))), "aa_seqs.fasta");
```

Appendix 3: Generating random pairs of sequences (script3.py)

```
#!/usr/bin/python

#Concatenate interacting pairs - encoded protein sequences in svm input format

import sys

sys.stdout = file(sys.argv[3],"a++")

parasite_input = file(sys.argv[1],"r")
partner_prots = file(sys.argv[2],"r")

if len(sys.argv) < 2:
        print "No files specified"
        sys.exit()
else:
        par_prot = parasite_input.readlines()[1]
        lines = partner_prots.readlines()[1:]
        for i, line in enumerate(lines):
                print par_prot.rstrip() + " " + line.lstrip(),
```

Appendix 4: Retrieving sequences from Genbank (script4.pl)

```
#!/usr/bin/perl -w

use strict;

use Bio::DB::GenBank;

open(my $file1, ">>", "/home/tseganesh/JKUAT_thesis_2014/set_A.fasta") or die $!;
open(my $file2, "<", "/home/tseganesh/JKUAT_thesis_2014/aaa.txt") or die $!;

my @array;
while(my $protA = <$file2>) {
    chomp $protA;
    push @array, $protA;
}

my $db_obj = Bio::DB::GenBank->new;

foreach my $acc(@array) {
    my $seq_obj = $db_obj->get_Seq_by_version($acc);
    print $file1 ">" . "$acc\n";
    print $file1 $seq_obj->seq;
    print $file1 "\n";
}

close $file1;
close $file2;
```

Appendix 5: Generating protein feature vectors (script5.py)

```
#!/usr/bin/python

#Concatenate interacting pairs - encoded protein sequences in svm input format

import sys

sys.stdout = file(sys.argv[3],"a++")

parasite_input = file(sys.argv[1],"r")
partner_prots = file(sys.argv[2],"r")

if len(sys.argv) < 2:
        print "No files specified"
        sys.exit()
else:
        par_prot = parasite_input.readlines()[1]
        lines = partner_prots.readlines()[1:]
        for i, line in enumerate(lines):
                print par_prot.rstrip() + " " + line.lstrip(),
```

Appendix 6: Preparation of input files into SVM format (script6.py)

```
#!/usr/bin/python

#put the encoded protein sequences in svm input format, indexed feature vectors

import sys
sys.stdout = file(sys.argv[3],"a") #output / indexed file
print "#", sys.argv[2] #this is the parasite protein ID that interacts with these host proteins

partner_prots = file(sys.argv[1],"r") #input / indexed file

for line in partner_prots:
    line = line.split(",")
    m = 1018; k = 1;
    while m < 2034:
        while k < 1018:
            print (str(m) + ":" + line[k]),
            m += 1
            k += 1
        else:
            break
    else:
        break
```