THE UNIVERSITY OF QUEENSLAND

AUSTRALIA

Comparative genomics of schistosomes: Identifying parasites specific genes, drug and vaccine targets

Md Shihab Hasan

Master of Science in Bioinformatics

Bachelor of Pharmacy (Honours)

A thesis submitted for the degree of Doctor of Philosophy at

The University of Queensland in 2017

Faculty of Medicine

Abstract

Schistosomiasis is a chronic disease, caused by Schistosoma species, affecting 200 million people worldwide and causing at least 300,000 deaths annually. Currently, no vaccines are available and Praziquantel is the standard anti-schistosomiasis drug. Praziquantel disrupts the tegument of adult worms, but not juvenile parasites and it does not prevent reinfection. Praziquantel resistance is rare, but repeated treatment in the field and laboratory manipulation has increased parasitic resistance. Therefore, it is necessary to develop a vaccine that induces long-term immunity to schistosomiasis with the final goal of complete elimination. Driven by the need to improve disease treatment and prevention, the genomes of three human *Schistosoma* species have recently become publicly available (*S. mansoni, S. japonicum* Chinese strain and *S. haematobium*). The principal goal of the PhD research project is to employ machine learning and Bioinformatics methods to identify novel vaccine and drug targets against the human-infecting *Schistosoma* parasites from genome sequence information.

In the first study, schistosome specific machine learning classifiers were developed for surface proteins and secretory peptides. Schistosome surface proteins, especially those expressed in tegument, represents the interface between host and parasite and its molecules are responsible for essential functions to parasite survival. Also, large number of proteins secreted by schistosomes are important for their survival in their hosts and infection. Knowledge of schistosome surface and secreted proteins is essential for understanding parasite host interaction and finding new candidate targets for vaccines and drugs or developing novel diagnostic methods. The web application SchistoProt has been developed, a schistosome specific classifier, for identifying schistosome specific surface proteins and secretory peptides that might be potential drug and vaccine targets.

In the second study, a machine learning prediction tool is developed to predict schistosome specific immunoreactive peptides. The sequence properties of immunoreactive *Schistosoma* proteins have been determined and compared the significant sequence features of immunoreactive proteins and non-immunoreactive proteins of *Schistosoma* species. The SchistoTarget web application, for the *in silico* identification of *Schistosoma* immunoreactive proteins has been developed. SchistoTarget uses supervised machine learning methods and significant differential features distribution between immunoreactive and non-immunoreactive peptides.

In the third study, a comparative analysis of the publicly available *Schistosoma* genomes *S. mansoni*, *S. Japonicum, S. haematobium*, the newly sequenced *Schistosoma bovis*

genome and the non-parasitic, free-living flatworm *Schmidtea mediterranea* reveals the interesting candidate genes for vaccine targets. Selected genes from this study have been annotated as surface or secretory proteins using the developed web applications from previous two studies. Further, using Gene Ontology and Swiss-Prot annotations, 20 putative vaccine and drug targets have been identified to be biologically validated by wet laboratory experiments in animals and then clinically.

The *in silico* comparative genomics analysis approach for identifying new drug and vaccine candidates represents a valuable resource for the *Schistosoma* research community. The protocol developed in this PhD research project can be used as a blueprint for other important parasitic diseases including malaria.

Declaration by author

This thesis *is composed of my original work, and contains* no material previously published or written by another person except where due reference has been made in the text. I have clearly stated the contribution by others to jointly-authored works that I have included in my thesis.

I have clearly stated the contribution of others to my thesis as a whole, including statistical assistance, survey design, data analysis, significant technical procedures, professional editorial advice, and any other original research work used or reported in my thesis. The content of my thesis is the result of work I have carried out since the commencement of my research higher degree candidature and does not include a substantial part of work that has been submitted *to qualify for the award of any* other degree or diploma in any university or other tertiary institution. I have clearly stated which parts of my thesis, if any, have been submitted to qualify for another award.

I acknowledge that an electronic copy of my thesis must be lodged with the University Library and, subject to the policy and procedures of The University of Queensland, the thesis be made available for research and study in accordance with the Copyright Act 1968 unless a period of embargo has been approved by the Dean of the Graduate School.

I acknowledge that copyright of all material contained in my thesis resides with the copyright holder(s) of that material. Where appropriate I have obtained copyright permission from the copyright holder to reproduce material in this thesis.

Publications during candidature

Peer-reviewed journals

1) Carla Proietti, Martha Zakrzewski, Thomas Watkins, Bernard Berger, Shihab Hasan, Champa Ratnatunga, Marie-Jo Brion, Peter Crompton, John Miles, Denise Doolan, Lutz Krause. 2016. Mining, visualizing and comparing multidimensional biomolecular data using the Genomics Data Miner (GMine) Web-Server. *Scientific Reports*. 6. DOI:10.1038/srep38178

2) Martha Zakrzewski, Carla Proietti, Jonathan J. Ellis, Shihab Hasan, Marie-Jo Brion, Bernard Berger, Lutz Krause. 2016. Calypso: A User-Friendly Web-Server for Mining and Visualizing Microbiome-Environment Interactions. *Bioinformatics*. DOI: 10.1093/bioinformatics/btw725

3) Lilian Pukk, Freed Ahmad, Shihab Hasan, Veljo Kisand, Riho Gross, Anti Vasemägi. 2015. Less is more: extreme genome complexity reduction with ddRAD using Ion Torrent semiconductor technology. *Molecular Ecology Resources*. 15(5). DOI: 10.1111/1755-0998.12392.

Conference Abstracts

1) Shihab Hasan, Martha Zakrzewski, Carla Proietti, Denise Doolan, Donald P. McManus, Lutz Krause. 2016. Bioinformatics approach for identifying schistosome immunoreactive proteins. International Congress for Tropical Medicine and Malaria 2016, 18 - 22 September 2016, Brisbane, Australia.

2) Shihab Hasan, Martha Zakrzewski, Don McManus, Lutz Krause. 2015. Vaccine targets from genomic sequence information – lessons learnt from schistosomes. Big Biology and Bioinformatics | B3 2015 Symposium. 23–24 November 2015, QUT Gardens Point, Brisbane, Australia.

3) Md Shihab Hasan, Martha Zakrzewski, Don McManus, Lutz Krause. 2015. Finding a Needle in a Haystack: Bioinformatic Approach for Identifying Vaccine Targets by Comparative Studies of Schistosoma Genomes & Proteomes. Great Lakes Bioinformatics Conference. 18-20 May 2015, Purdue University, West Lafayette, USA.

4) Shihab Hasan, Martha Zakrzewski, Don McManus, Lutz Krause. 2014. Bioinformatic Approach for Identifying Schistosome Specific Surface and Secretory proteins. Big Biology and Bioinformatics | B3 2014 Symposium. 24–26 November 2014, QUT Gardens Point, Brisbane, Australia.

Publications included in this thesis

None.

Contributions by others to the thesis

My principal advisor Associate Prof. Lutz Krause and associate advisors Dr. Martha Zakrzewski and Prof. Donald P. McManus contributed to the conception and design of this research, advised on methods and analyses, and provided critical comments on the thesis.

Statement of parts of the thesis submitted to qualify for the award of another degree

None.

Acknowledgements

I dedicate this thesis to my father Mr. Lutfor Rahman who has passed away during the PhD candidature.

The work presented in this thesis would not have been possible without the advice, guidance, encouragement and support of my principal advisor Associate Professor Lutz Krause, and co-supervisors Dr. Martha Zakrzewski and Prof. Donald P. McManus. All your guidance, advice and skills you taught me are invaluable and will be helpful in my future career. I would like to express my sincere gratitude to all of you.

In addition, many thanks to Prof. Denise Doolan and Dr. Katja Fischer for providing the advice given throughout this degree.

The completion of this thesis would not have been possible without scholarship support from the University of Queensland (University of Queensland International Scholarship and Research Higher Degree Top-Up Scholarship) and QIMR Berghofer Medical Research Institute (QIMR Berghofer International Scholarship and Top-up scholarship).

Finally, and most important, my infinite gratitude to my family and friends for all their support during always.

Keywords

schistosomiasis, bioinformatics, machine learning, vaccine targets, surface proteins, secretory proteins, immnunoreactive proteins, comparative genomics, web server

Australian and New Zealand Standard Research Classifications (ANZSRC)

ANZSRC code: 060102, Bioinformatics, 30%

ANZSRC code: 060408, Genomics, 30%

ANZSRC code: 080301, Bioinformatics Software, 40%

Fields of Research (FoR) Classification

FoR code: 0601, Biochemistry and Cell Biology, 30%

FoR code: 0604, Genetics, 30%

FoR code: 0803, Computer Software, 40%

Table of Contents

List of Figures

List of Tables

List of Supplementary Tables

List of Abbreviations

Abbreviation	Definition
ATPase	Adenosine 5'-triphosphatase
AUC	Area Under the curve
BNB	Bernoulli Naive Bayes
Bp	Base pair
C3	Complement component 3
C4	Complement component 4
CD4	Cluster of differentiation 4
CD44	Cluster of Differentiation 44
CSS	Cascading style sheets
CT-SOD	Chaetomium thermophilum Superoxide Dismutases
ECL	Enhanced chemiluminescence
FDR	False discovery rate
FN	False Negative
FP	False Positive
GBM	Gradient Boosting Machine
GNB	Gaussian Naive Bayes
GO	Gene Ontology
GRAVY	Grand average of hydropathy
HMMs	Hidden Markov models
HSPs	High-scoring Segment Pairs
HTML	Hypertext Markup Language
IFNγ	Interferon gamma
IgE	Immunoglobulin E
IgG	Immunoglobulin G
IL7	Interleukin 7
LDA	Linear Discriminant Analysis
MgATP	Magnesium Adenosine 5'-triphosphate
MHC	Major histocompatibility complex
MLP	Multi-layer Perceptron
mRNA	Messenger RNA
mTP	Mitochondrial targeting peptide

NCBI	National Center for Biotechnology Information
NTDs	Neglected tropical diseases
PANTHER	Protein annotation through evolutionary relationship
PSSMs	Position weight matrices
PZQ	Praziquantel
QDA	Quadratic Discriminant Analysis
RBF	Radial basis function
RBH	Reciprocal Best Hits
ROC	Receiver Operating Characteristic
RTS	rapid translation system
SGD	Stochastic gradient descent
Sjp	*Schistosoma japonicum*
Sm	*Schistosoma mansoni*
Smteg	*Schistosoma mansoni* schistosomula tegument
SP	Secretory pathway signal peptide
SVM	Support Vector Machine
TGF-β	Transforming growth factor beta
TN	True Negative
TNF-α	Tumor necrosis factor alpha
TP	True Positive
TSP2	Thrombospondin 2

Chapter 1 Introduction

1.1 Schistosomiasis

Infections by blood flukes (schistosomes) cause highly significant human diseases and are a major health concern in the Asia Pacific Region and Africa. Schistosomiasis is a chronic disease caused by Schistosoma species. It is considered by the World Health Organization as the second most socioeconomically devastating and second most common parasitic disease affecting 200 million people worldwide[1,2] and causing at least 300,000 deaths annually[3]. No vaccines are available and treatment relies mainly on one drug, praziquantel[2]. Eight Schistosoma species infect humans: *S. mansoni*, *S. haematobium, S. japonicum, S. mekongi, S. malayensis, S. mattheei, S. guineeensis, and S. intercalatum*[4] (Figure 1.1). Recent applications of next-generation sequencing technologies and bioinformatic tools for large-scale investigations explore the systems biology of the organisms[5]. The genome sequences provide a unique resource for studying the evolution of schistosomes, to identify genes important for host-parasite interactions and to discover novel drug and vaccine targets.

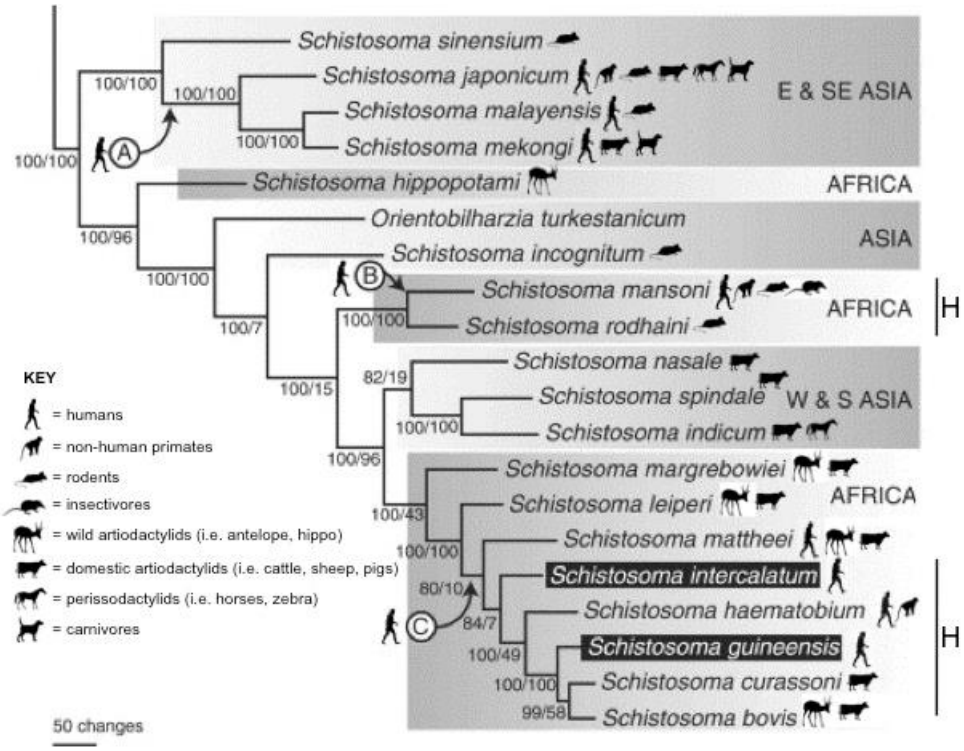


Figure 1.1 Phylogeny of Schistosoma.

(Figure adapted[4,6])

This chapter will first introduce background knowledge of the schistosome lifecycle, schistosomiasis treatment, infection and resistance, available schistosome genome data and potential vaccine targets.

1.2 Schistosome Structure and Lifecycle

Adult schistosomes are white or greyish worms with a cylindrical body of 7–20 mm in length. The body has two terminal suckers, a blind digestive tract, reproductive organs and a complex tegument. The tegument consists of a single, contiguous, double-bilayered membrane, which covers the entire worm[7]. Schistosomes are exposed to diverse environmental conditions during their life cycle. Unlike other trematodes, schistosomes have separate sexes. They change from an asexual form in the intermediate hosts such as snails. Then they change to a sexual form in the vascular system of the definitive host such as human[8].

The male's body holds the longer and thinner female by forming a groove or gynaecophoric channel (Figure 1.2). The adult schistosomes live within the perivesical (*Schistosoma haematobium*) or mesenteric (other species) venous plexus as permanently embraced couples. Schistosomes feed on blood and globulins through anaerobic glycolysis, and the debris released in the host's blood[9].

Figure 1.2 Paired adult *S. mansoni* worms.

The darker female lying within the gynacophoric canal of the larger male worm.
(From Schistosomiasis Research Group, http://www.path.cam.ac.uk/~schisto)

Schistosomes require an intermediate aquatic snail host in their complex life cycle (Figure 1.3). The sporocyst (snail stage) reproduces asexually, producing cercariae that are

constantly shed into the aquatic environment. These cercariae find the definite host (human or animal) in contact with the water and penetrate the skin. The larva transforms into a schistosomulum in the skin, that adapts its surface membrane, the tegument, for parasitism. The schistosomulum passes from the skin, through the lymphatic system and blood into the lungs and via the blood to the liver. Within the liver the parasites form sexual pairs and develop into adult worms. The adult worm pairs live and lay eggs in the vessels surrounding the intestine or urinary system. Mature eggs can penetrate host membranes such as rectal veins or the intestinal wall. Eggs are released from the host body and all eggs that come into contact with water hatch into miracidia and the cycle starts again[10].

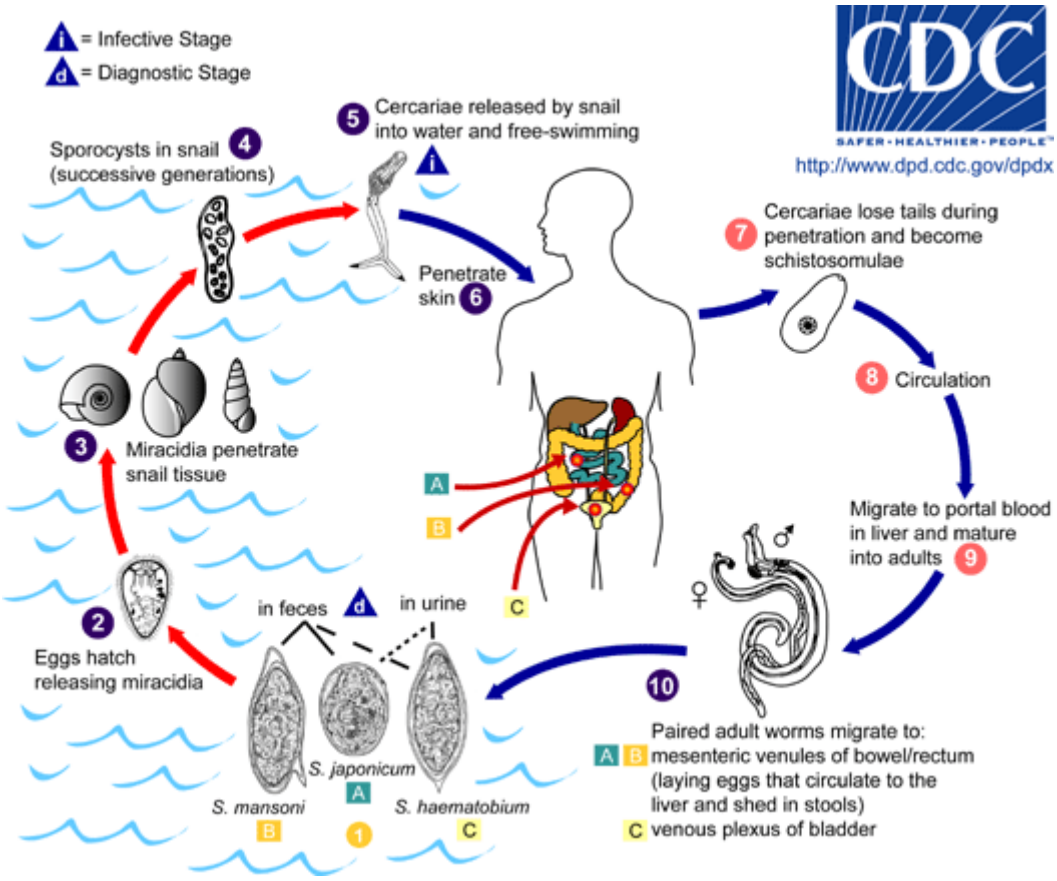


Figure 1.3 The schistosome life-cycle.

(From Schistosomiasis Image Library, www.dpd.cdc.gov)

1.3 Schistosomiasis Treatment

Currently, Praziquantel (PZQ) is the standard anti-schistosomiasis drug[11]. Praziquantel disrupts the tegument of adult worms, but not juvenile parasites[9] and it does not prevent reinfection[12]. Praziquantel resistance is rare, but repeated treatment in the field and laboratory manipulation has increased parasitic resistance[11,13,14]. Therefore, it is necessary

to develop a vaccine that induces long-term immunity to schistosomiasis with the final goal of complete elimination[15].

One vaccine candidate is the rSh28GST antigen from S. haematobium, currently in phase I clinical trial and shown to be safe and immunogenic[16]. Other vaccine candidates are the Sm14, Sm29, Sm-TSP1 and Sm-TSP2 antigens from S. mansoni, currently in pre-clinical and clinical development[17-19].

1.4 Available *Schistosoma* Genomes

Driven by the need to improve disease treatment and prevention, the genomes of three human *Schistosoma* species have recently become publicly available (*S. mansoni*[20], *S. japonicum* Chinese strain[21] and *S. haematobium*[22]). Comparison of these three genomes shows similar genome size, number of proteins and similar GC content and percentage of repetitive elements (Table 1.1).

Table 1.1 Comparison of three publicly available Schistosoma genomes.

Genomic features	*Schistosoma mansoni*	*Schistosoma japonicum*	*Schistosoma haematobium*
Estimate of genome size (Mb)	381	403	385
Chromosome number (2*n*)	8	8	8
Total number of base pairs within assembled contigs	374,944,597	369,039,322	361,903,089
N50 contig (length (bp); total number >500 bp)	16,320; *n* = 50,292	6,121; *n* = 95,265	21,744; *n* = 36,826
Total number of base pairs within assembled scaffolds	381,096,674	402,705,545	385,110,549
N50 scaffold (length (bp); total number >1,000 bp in length)	832,5415; *n* = 19,022	176,869; *n* = 25,048	306,738; *n* = 7,475
Proportion of genome that is coding (%)	4.72	4.32	4.43
Number of putative coding genes	13,184	13,469	13,073
Total GC content (%)	34.7	33.5	34.3
Repeat rate (%)	45	40.1	47.2

(Source[22])

1.5 *Schistosoma* Surface Proteins and Secretory Peptides

The body of the adult worm is covered by a complex multilaminate surface, the tegument, which enables schistosomes to survive in the hostile host environment for decades. Schistosomes also display effective strategies to evade the host immune responses[23,24]. A large number of proteins are excreted or secreted by schistosomas from their surface. These excretory proteins are important for their survival in their hosts. These proteins can stimulate the innate immune system and modulate various host immune responses when exposed to host tissues. Thus, schistosomas evade the host immune defense and become resistance to antibody-dependent cellular cytotoxicity and oxidative stress[25,26]. Identification of surface proteins and secreted peptides is important for both understanding parasite host interaction and finding new candidate vaccines targets[27].

1.6 Host-Parasite Interaction

Significantly acclimated to parasitic life, schistosomes can live for a long time or decades even in a hostile environment as the circulatory system from vertebrate host[28]. The parasite has a close contact with circulating elements of the immune system[29]. In this effective host-parasite relationship, the host immune system plays an important role in both parasite development and elimination. CD4+ cells, hormones, and cytokines as TNF-α, TGF-β, and IL-7 produced by the host, appear to aid the parasite development (Figure 1.4), suggesting that schistosomes could accept host hormone signals for cell proliferation, development, mating, and reproduction while CD4+ cells, B cells, IFN-γ, and TNF-α have been implicated in parasite elimination in the irradiated cercariae vaccine model[30,31].

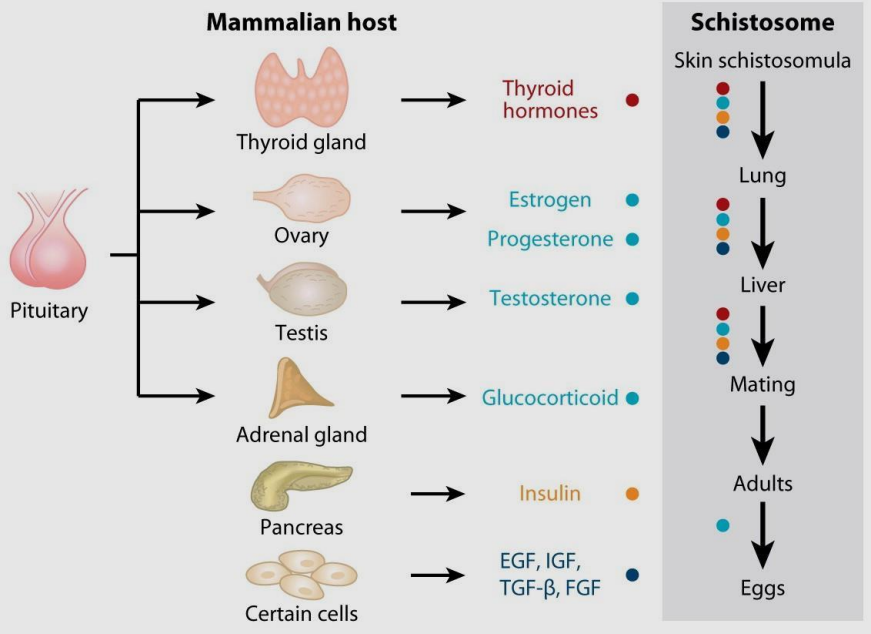


Figure 1.4 A theoretical model of the impact of a few hormones and development elements from a mammalian host on certain developmental stages of schistosomes.

(Source[30])

1.7 Antigens for Vaccine Development

The majority of the studies that planned to recognise membrane proteins in parasite tegument were performed in adult worms[23,32]. Schistosomula is the significant focus for host immunity. Protective antigens are found in *S. mansoni* schistosomula tegument (Smteg) since mice inoculation with Smteg formulated with Freunds' adjuvant[33]. The characterization of these defensive antigens is being performed utilising immune-proteomics analysis and genome databases to distinguish candidates to be utilised in a vaccine formulation against schistosomiasis[31,34]. The results observed in these preclinical trials using tegument proteins are summarised in Table 1.2.

Table 1.2 Schistosome tegument protein evaluated as vaccine candidates in preclinical studies.

(Table adapted[31,35])

Protein	Vaccine type	Protection level	Egg reduction	Bioinformatic tool used in antigen selection
Sm 21.7	Recombinant protein	41%–70%	Not determined	Not determined

Sm 21.7	DNA vaccine	41.5%	62% (liver) 67% (intestine)	Not determined
Cu/Zn superoxide dismutase	DNA vaccine	44%–60%	Not determined	Not determined
Sm TSP2	Recombinant protein	57%	64% (liver) 65% (feces)	BLAST
Sm29	Recombinant protein	51%	60% (intestine)	InterProScan, SignalIP 3.0, Signal IP Neural, NetNGlyc 1.0, BLAST, WolfpSORT, SOSUI, Compute pI/Mw tool
ECL (200 kDa protein)	DNA vaccine	38.1%	Not determined	Not determined
Sm 22.6	Recombinant protein	34.5%	Not determined	BLAST
Sm TSP 1	Recombinant protein	34%	52% (liver) 69% (intestine)	BLAST
Sm 21.7	Recombinant protein	41%–70%	Not determined	Not determined
Sm p80	DNA vaccine	Not determined	Not determined	Not determined
Sm14e	Recombinant protein	Not determined	Not determined	Not determined
CT-SOD	DNA vaccine	Not determined	Not determined	Not determined

The immunomic screening of pathogens using protein microarrays for antigen discovery has progressed rapidly. The first protein microarray for schistosomes has been constructed for identification of valuable immunogens that could be developed as marketable vaccines[15]. Recently, another proteome microarray of *S. mansoni* proteins was produced[36].

1.8 Hypothesis

During the past decade, *Schistosoma* researchers targeted surface and secretory proteins for vaccine development. But, a very limited number of surface and secretory antigens were explored[23,24,37-39]. Recently, three human-infecting Schistosoma genomes become publicly available. Also, immunomics approach using schistosome protein microarray provide useful resource on antibody responses of the antigens. The genomics and immunomics datasets provide a unique foundation for an innovative approach to identify novel drug and vaccine targets but it has not been obvious how to identify the important protective schistosome antigens from genomic-based information.

The hypothesis is that bioinformatics approach leads to identifying putative drug and vaccine targets against schistosomiasis using *Schistosoma* genomic-based information. In this PhD project, I have developed an integrative bioinformatics pipeline to identify putative drug and vaccine targets against schistosomiasis from protein sequences information.

1.9 Research Aims

The principal goals of the proposed project are to employ Bioinformatics methods to identify novel vaccine and drug targets against the human parasites *Schistosoma spp*.

1.9.1 Aim 1: Identify schistosome-specific Surface Proteins and Secreted Peptides

Develop schistosome-specific machine-learning classifier for the identification of surface proteins and secreted peptides. Apply newly developed classifier to in all in-house and publicly available schistosome genomes.

1.9.2 Aim 2: Identify *Schistosoma* immunoreactive proteins

Develop schistosome-specific machine-learning classifier for the identification of immunoreactive proteins using *Scistosoma* protein microarray data.

1.9.3 Aim 3: Identify putative vaccine targets against schistosomiasis

Develop an integrative bioinformatics pipeline to identify putative vaccine targets against schistosomiasis by comparative analysis of available genomic-based information.

1.10 Thesis outline

In chapter 1, literatures are reviewed to extract background information required for schistosome proteins characteristics to select the potential antigens as drug and vaccine targets. The development of methods and associated tool, SchistoProt, to identify schistosome-specific surface proteins and secretory peptides are described in chapter 2. Chapter 3 describes a machine learning approach, the SchistoTarget web server, to identify *Shistosoma* proteins immunoreactivity using protein microarray data. A comparative analysis of Schistosoma genomes and an integrative bioinformatics pipeline to identify putative vaccine targets against schistosomiasis has been described in Chapter 4. In this chapter, I have shown how potential antigens can be selected by comparing several parasite genomes and using the tools developed in chapters 2 and 3 and other available annotations. A general discussion on the research outcomes from the PhD project and future direction are provided in chapter 5.

Chapter 2 Identifying Schistosome-Specific Surface Proteins and Secretory Peptides

2.1 Foreword

This chapter describes a supervised machine learning based approach used for identifying *Schistosoma*-specific surface proteins and secreted peptides. A machine learning based web server, SchistoProt, has been developed to classify *Schistosoma* protein sequences into surface/non-surface proteins and secretory/non-secretory peptides. The methods, prediction accuracy, usage and architecture of SchistoProt have been depicted in this chapter.

2.2 Abstract

Schistosomiasis is a debilitating chronic disease caused by *Schistosoma* parasitic worms. It is considered by the World Health Organization as the second most devastating parasitic disease, with a strong need for vaccine development. Knowledge of schistosome surface and secreted proteins is essential for understanding parasite-host interactions, for studying anti-*Schistosoma* protective immunity, for finding new candidate vaccine targets, and for developing novel diagnostic methods.

SchistoProt, a web-based classifier for the *in silico* identification of schistosome surface proteins and secreted peptides, have been inroduced. The classifier is highly accurate and fast, and allows the analysis of large whole-proteome datasets. Positive training sets (known surface and non-secretory proteins) were extracted from the literature and the NCBI non-redundant protein database. A negative training set was compiled from nuclear and histone related proteins. SchistoProt provides a user-friendly web-interface and results are presented in interactive tables and figures. On an independent test-set of 400 *Schistosoma* proteins, SchistoProt achieved a sensitivity of 85% and specificity of 81% for surface proteins and a sensitivity of 92% and specificity of 93% for secretory peptides. The software showed significantly increased prediction accuracy compared with existing tools. SchistoProt is implemented in Python and the web-server is freely accessible at http://schistoprot.bioapps.org. Source code and documentation are available from https://github.com/shihabhasan/schistoprot.

SchistoProt is an easy-to-use, highly accurate and fast web-server for the *in silico* identification of *Schistosoma* surface proteins and secreted peptides. The software has been optimized for large datasets and enables whole-proteome analysis. SchistoProt can assist rational vaccine design by facilitating the rapid prioritization of candidate vaccine targets. The software also identifies proteins potentially important for parasite-host interaction and therefore enables researchers to gain new insights into the molecular mechanisms of *Schistosoma* infection.

2.3 Introduction

Schistosomiasis is an infectious disease caused by parasitic *Schistosoma* worms[1,2]. More than 700 million individuals are at risk of acquiring schistosomiasis in more than 70 countries. It is considered by the World Health Organization as the second most socioeconomically devastating and second most common parasitic disease after malaria[3]. Chemotherapy via praziquantel is an effective treatment, but mass treatment does not prevent reinfection and there is an increasing concern of the development of drug resistance. The development of vaccines that induce long-term immunity therefore remain the most potentially effective means for controlling schistosomiasis. However, despite the poor containment of the disease and devastating medical and economic impact, no *Schistosoma* vaccines are available and we are just starting to understand the molecular mechanisms of host infection, host-parasite interaction and anti-schistosome protective immunity.

Driven by the need to improve disease treatment and prevention, the genomes of three human *Schistosoma* species have recently become publicly available[20-22]. The surface of larval and adult schistosomes, the tegument, represents the host-parasite interface and proteins expressed in the tegument are responsible for essential functions for parasite survival in the host[23]. The tegument includes a single multinucleated cytoplasmic layer, which is linked to underlying nucleated cell bodies by cytoplasmic connections, that covers the entire worm[40]. Proteomic analysis of *S. mansoni* surface proteins showed the presence of enolase, an enzyme involved in energy metabolism, and structural molecules such as calcium ATPase which can inhibit platelet activation[37,41]. The leukocyte marker CD44, host complement proteins C3 and C4, and the membrane protease calpain have also been identified in *S. mansoni* surface proteins[42].

Proteins secreted by schistosomes are also essential for infection, e.g. by modulating host immune responses[43]. A number of endo- and exo-peptidases, trypsin-type serine

peptidase(s), and metallo-peptidases, have been revealed in the secretory proteins of *S. mansoni*[44]. These secretory proteins can stimulate the innate immune system and modulate various host immune responses which help the parasite evade immune defence mechanism when exposed to the host environment[25]. Knowledge of *Schistosoma* surface proteins and secreted peptides is therefore essential for improving our understanding of host-parasite interaction and for rational vaccine design.

Existing approaches for the *in silico* prediction of surface and secretory proteins use hidden Markov models (HMMs), Bayesian networks, neural networks or position weight matrices (PSSMs)[45-49]. Two of the first and most widely used tools for the prediction of transmembrane proteins and signal peptides are TMHMM and SignalP, respectively. TMHMM employs a HMM to represent the different sequence regions of transmembrane helices[47]. SignalP relies on a combination of several artificial neural networks to predict the presence and location of signal peptide cleavage sites[48]. The recently developed combined classifier Phobius uses a HMM to model the sequence properties of both signal peptides and transmembrane proteins[45]. Philius predicts transmembrane topology and signal peptides using dynamic Bayesian networks[46]. PrediSi is based on position weight matrices to identify signal peptides and their cleavage positions and has been developed for the rapid analysis of whole-proteome datasets[49]. All of these approaches are general classifiers which perform well for a wide range of bacterial and eukaryotic species, but show a modest prediction accuracy for *Schistosoma* species[27]. Liao *et al*. have recently demonstrated that a *Schistosoma*-specific classifier can significantly increase the prediction accuracy for identifying secreted proteins[27]. However, at present, no genus-specific classifier is available for predicting *Schistosoma* secretory peptides and surface proteins, which would be invaluable for improving our knowledge of *Schistosoma* host-parasite interactions, parasite pathogenesis and anti-schistosomiasis protective immunity. Such a tool will further assist researchers in discovering urgently needed anti-schistosomal vaccines.

To address this need, the SchistoProt web server, a genus-specific, highly accurate machine learning classifier for identifying *Schistosoma* surface proteins and secretory peptides, have been developed. The server relies on 3 supervised machine learning techniques, evaluates a wide range of different sequence properties for classification, and is freely available at http://schistoprot.bioapps.org.

2.4 Methods

2.4.1 Supervised machine learning classification

Machine learning is learning is a field of computer science that provides systems (computers) the ability to automatically learn and improve from experience without being explicitly programmed. Supervised machine learning classification is the machine learning task of identifying to which of a set of classes a new observation belongs from labelled training data which consists of a set of training samples whose classes are known[50].

2.4.2 Training set

To classify *Schistosoma* proteins into surface/non-surface and secreted/non-secreted classes SchistoProt uses 3 different supervised machine learning techniques. First, the 3 classifiers had to be trained on a so called positive and negative training set to learn the specific properties of each class. The positive training set of tegument/surface proteins consisted of 414 sequences, which have been extracted from the published literature[7,15,24,36,39,51,52] and from experimentally validated (not computationally predicted) protein sequences from the NCBI non-redundant protein database. For the negative training set (non-surface proteins), 435 nuclear, histone and mitochondrial related proteins were collected from the literature[24,27] and from validated sequences from the NCBI non-redundant protein database. As a positive training set for secreted proteins, a total of 375 proteins were collected from the literature[27,38,43,51] and from validated sequences from the NCBI non-redundant protein database. For the negative training set (non-secretory proteins) 746 nuclear and histone related proteins were collected from the literature[24,27] and from validated sequences from the NCBI non-redundant protein database. Only experimentally validated proteins were included in both datasets to obtain high-quality and reliable training sets. PISCES[53] was applied to remove proteins with sequence identity over 20% to reduce biases towards overrepresented proteins. A total of 249 surface proteins, 277 non-surface proteins, 205 secreted proteins and 258 non-secreted proteins remained in the final training sets. Seven proteins were present in both the surface and secretory positive training datasets.

2.4.3 Independent test set T400

To evaluate the classification accuracy of SchistoProt, an independent test set (named T400) comprising 400 *Schistosoma* proteins has been compiled. 100 surface proteins, 100

secreted proteins, 100 non-surface and 100 non-secreted proteins from the NCBI non-redundant proteins database were randomly selected. Using sequence similarity comparisons, it is ensured that none of the 400 selected proteins was present in the training sets.

2.4.4 Features used for SchistoProt classification

Initially, 481 features from each protein were extracted. Out of these 481 features, 81 features represent sequence characteristics and structural and biochemical attributes (

Supplementary Table 2.1). The remaining 400 features are k-mers of 2 amino acid residues (2-mers) for 20 amino acids (Supplementary Table 2.2). 2-mers refer to all the possible subsequences of length 2 from a protein sequence. Features are extracted from each protein sequence using different available bioinformatics tools and newly developed Python scripts (Table 2.1). The features were approximately normally distributed which was tested based on the comparisons of mean and median values, and the shape of the data (Supplementary Table 2.3). The distribution of features between surface/non-surface proteins and between secreted/non-secreted peptides in the training set were compared by t-test. Only features that were significantly associated with one class ($p<0.01$, FDR<0.05) are used in SchistoProt (More Conservative mode) for the identification of surface proteins and secreted peptides, respectively.

Table 2.1 Tools used to extract protein features.

Tool	Purpose	URL
Pepstats	Calculation of statistics for proteins such as molecular weight, isoelectric point etc.	http://www.ebi.ac.uk/Tools/seqstats/emboss_pepstats/
Protparam	Computation of various physical and chemical parameters	http://web.expasy.org/protparam/
Garnier	Prediction of protein secondary structure	http://emboss.sourceforge.net/apps/release/6.2/emboss/apps/garnier.html
NetCGlyc	C-mannosylation sites	http://www.cbs.dtu.dk/services/NetCGlyc/
NetChop	Proteasomal cleavages (MHC ligands)	http://www.cbs.dtu.dk/services/NetChop/
NetNGlyc	N-linked glycosylation sites	http://www.cbs.dtu.dk/services/NetNGlyc/
ANCHOR	Prediction of Protein Binding Regions in Disordered Proteins	http://anchor.enzim.hu/
ProP	Arginine and lysine propeptide cleavage sites	http://www.cbs.dtu.dk/services/ProP/
TargetP	Prediction of the subcellular location of eukaryotic proteins	http://www.cbs.dtu.dk/services/TargetP/
BepiPred	Prediction of the location of linear B-cell epitopes	http://www.cbs.dtu.dk/services/BepiPred/
Class I Immunogenecity	Prediction of MHC Class I immunogenicity	http://tools.iedb.org/immunogenicity/

2.4.5 Feature scaling

The range of values of the different features included in SchistoProt varies widely. To ensure that each feature contributes approximately proportionately and, therefore, avoid biases introduced by features with greater numeric ranges[54], all features are scaled into the range of 0 to 1.

2.4.6 Selection of best performing machine-learning technique

Initially, 16 different machine learning techniques for all training sets were applied. Classifiers were run using the Scikit-learn (Version 0.18.1) library in Python[55] with optimized parameters (Supplementary Text 2.1). The following classifiers were used: (i) Gradient Boosting Machine (GBM); (ii) Support Vector Machine (SVM) with radial basis function (RBF SVM) kernel with C=16, gamma=0.01; (iii) K-Neighbors; (iv) Decision tree with max_depth=13; (v) Random forest with max_depth=13, n_estimators=13 for surface classifier and max_depth=15, n_estimators=15 for secretory classifier, and max_features=481; (vi) Ada boost; (vii) Gaussian Naive Bayes (GNB); (viii) Linear Discriminant Analysis (LDA); (ix) Quadratic Discriminant Analysis (QDA); (x) Ridge regression; (xi) Stochastic gradient descent (SGD); (xii) Perceptron; (xiii) Passive aggressive; (xiv) Bernoulli Naive Bayes (BNB); (xv) Nearest Centroid; and (xvi) Multi-layer Perceptron (MLP).

Each training sequence was represented by the corresponding feature vectors. The 16 classifiers were then evaluated by stratified k-fold (10-fold) cross-validation. Using stratified k-fold cross-validation, the folds were selected such that the mean response value was approximately equal in all folds[56]. In 10-fold cross-validation, 90% of the data were used for training and the remaining 10% for testing. The cross-validation process was repeated 10 times and the average predication accuracy calculated.

SchistoProt relies on the 3 supervised machine learning techniques which achieved the highest prediction accuracies in the 10-fold cross-validation. Gradient Boosting Machine (GBM), Random Forest and Bernoulli Naive Bayes (BNB) classifiers were selected for the classification of surface proteins. GBM, Ada Boost and BNB were selected for the classification of secretory proteins. The 3 machine-learning techniques are combined into a single classifier using a majority rule. A protein is assigned to positive class if it is predicted by at least 2 of the 3 classifiers as positive, otherwise, SchistoProt assigns the protein as negative class i.e., only one or no classifiers predict the protein as positive.

2.4.7 Performance evaluation

The classification accuracy of SchistoProt was evaluated by sensitivity, specificity and overall accuracy[57]. These measures are defined as: sensitivity = TP/(TP+FN), specificity = TN/(TN+FP), overall accuracy = (TP + TN) / (TP + TN + FP + FN), where TP (True Positive) and TN (True Negative) are the number of correctly predicted positive and negative proteins, respectively, and FP (False Positive) and FN (False Negative) are the

number of incorrectly predicted positive and negative proteins, respectively. Additionally, the discriminatory power of classifiers was evaluated by the Area Under the Receiver Operating Characteristic (ROC) curve (AUC).

2.5 Results

2.5.1 SchistoProt overview

SchistoProt uses 3 supervised machine-learning classifiers to discriminate between surface and non-surface proteins and between secreted and non-secreted peptides. Generated predictions are stored in a database which facilitates rapid reuse of results without rerunning the time-consuming classifiers. This saves considerable runtime if the same sequences are uploaded multiple times, e.g. by different users.

SchistoProt takes FASTA formatted sequence files or pasted protein sequences as input. If the proteins are already present in the database, the pre-computed results are returned. Otherwise SchistoProt extracts corresponding features from each query sequence using several available bioinformatics tools and newly developed Python scripts (Table 2.1). Features include sequence characteristics, biochemical attributes and structural properties (Supplementary Table 2.1). These features are scaled and used to discriminate between surface and non-surface proteins and between secreted and non-secreted peptides. SchistoProt combines 3 supervised machine learning techniques and classifies proteins based on a majority rule. The results of the classification are returned to the user and stored in the database for future reuse (Figure 2.1). Results are presented as interactive tables, charts and figures.

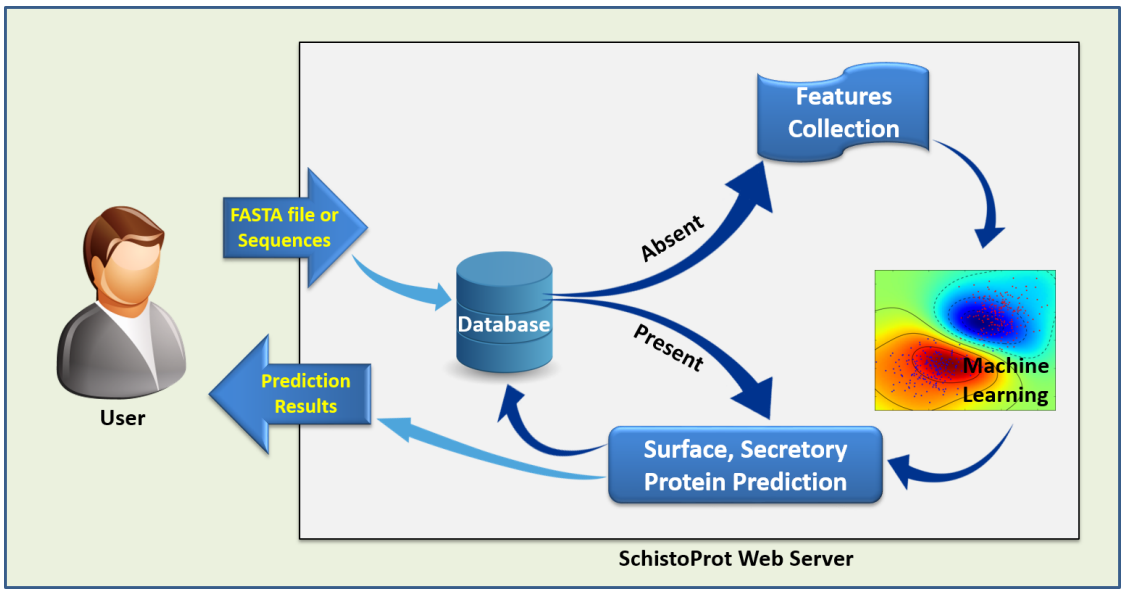


Figure 2.1 SchistoProt workflow.

Query sequences submitted by a user are compared to a database of pre-computed results. If a query sequence is present in the database, SchistoProt reports the pre-computed predictions; otherwise, a machine-learning classification is performed and the results are stored in the database for future reuse. Results are presented online as interactive tables, charts and figures.

2.5.2 Features significantly associated with surface proteins and secreted peptides

Associations between 81 biochemical and structural sequence features and surface proteins and secreted peptides have been examined using a t-test. Fifty-four features were significantly differentially distributed between surface and non-surface proteins ($p<0.01$, FDR<0.05) (Figure 2.2; Table 2.2). Surface proteins showed a higher frequency of lysine, isoleucines, secondary pathway signal peptides, and secondary helices. Surface proteins were also found to be more stable, aromatic and Class I immunogenic than non-surface proteins. Arginine and proline were underrepresented in surface proteins.

Fifty-seven features showed differential distribution between secretory and non-secretory proteins ($p<0.01$, FDR<0.05) (Figure 2.3; Table 2.3). Secretory proteins showed a higher frequency of grand average of hydropathy (GRAVY), non-polar moles, lysine and the hydrophobic amino acids glycine, valine, isoleucine, phenylalanine, methionine, and tryptophan. Secretory proteins were also more stable than non-secretory proteins. Secondary turns, polar moles and serines were higher in non-secretory proteins.

129 of 2-mers were significantly differentially distributed between surface and non-surface proteins ($p<0.01$, FDR<0.05) (Supplementary Table 2.4). 122 of 2-mers were significantly

differentially distributed between secretory and non-secretory proteins (p<0.01, FDR<0.05) (Supplementary Table 2.5).

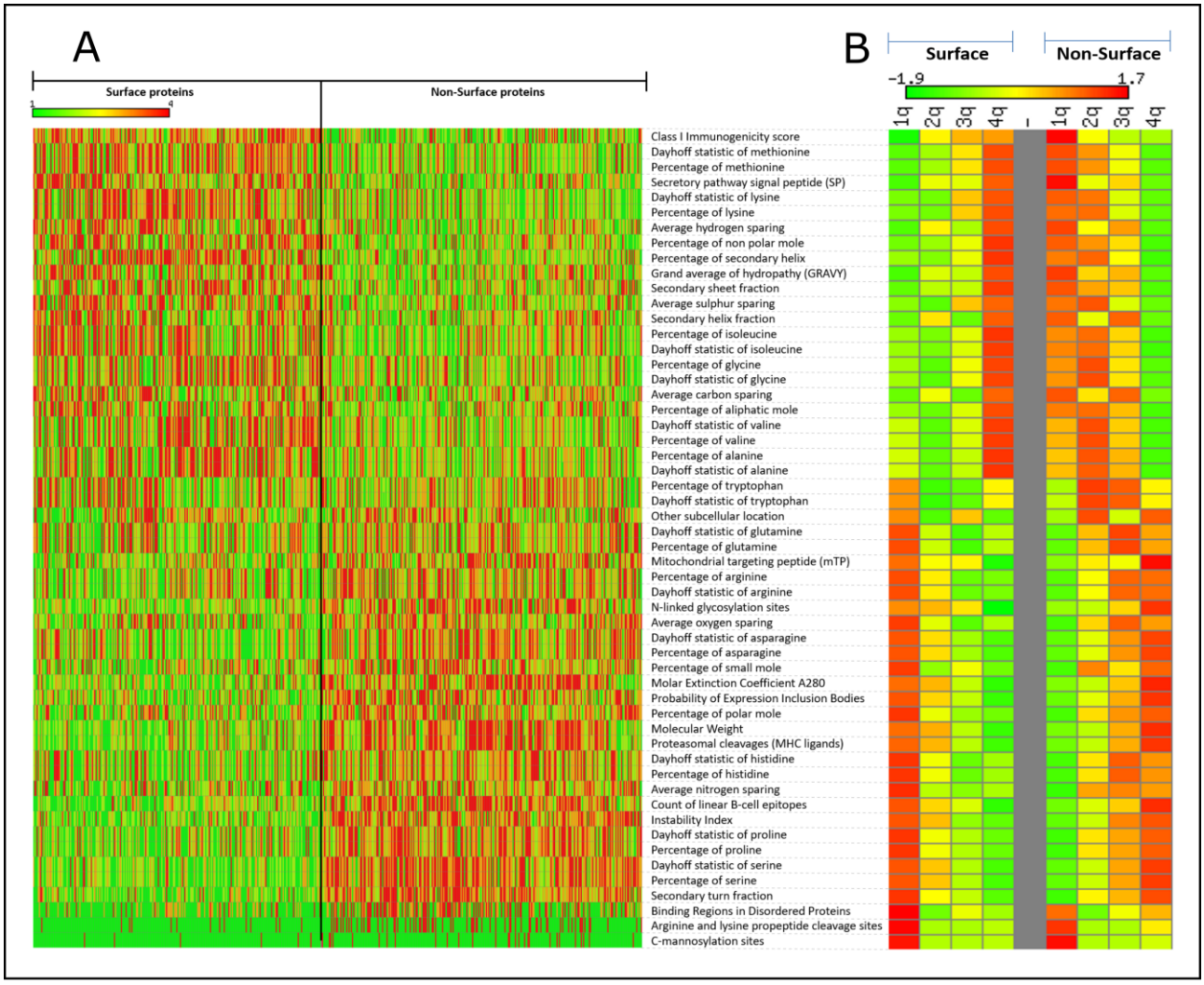


Figure 2.2 Features associated with surface proteins.

Means in the surface positive and surface negative training sets were compared by t-test. Shown are all features with p<0.01. (A) Heatmap of features significantly differentially distributed between surface and non-surface proteins. Columns represent each protein of the training set, rows represent features. (B) Quantiles distribution of significantly different features for surface and non-surface proteins. Values are depicted in color code, ranging from green (low) to red (high).

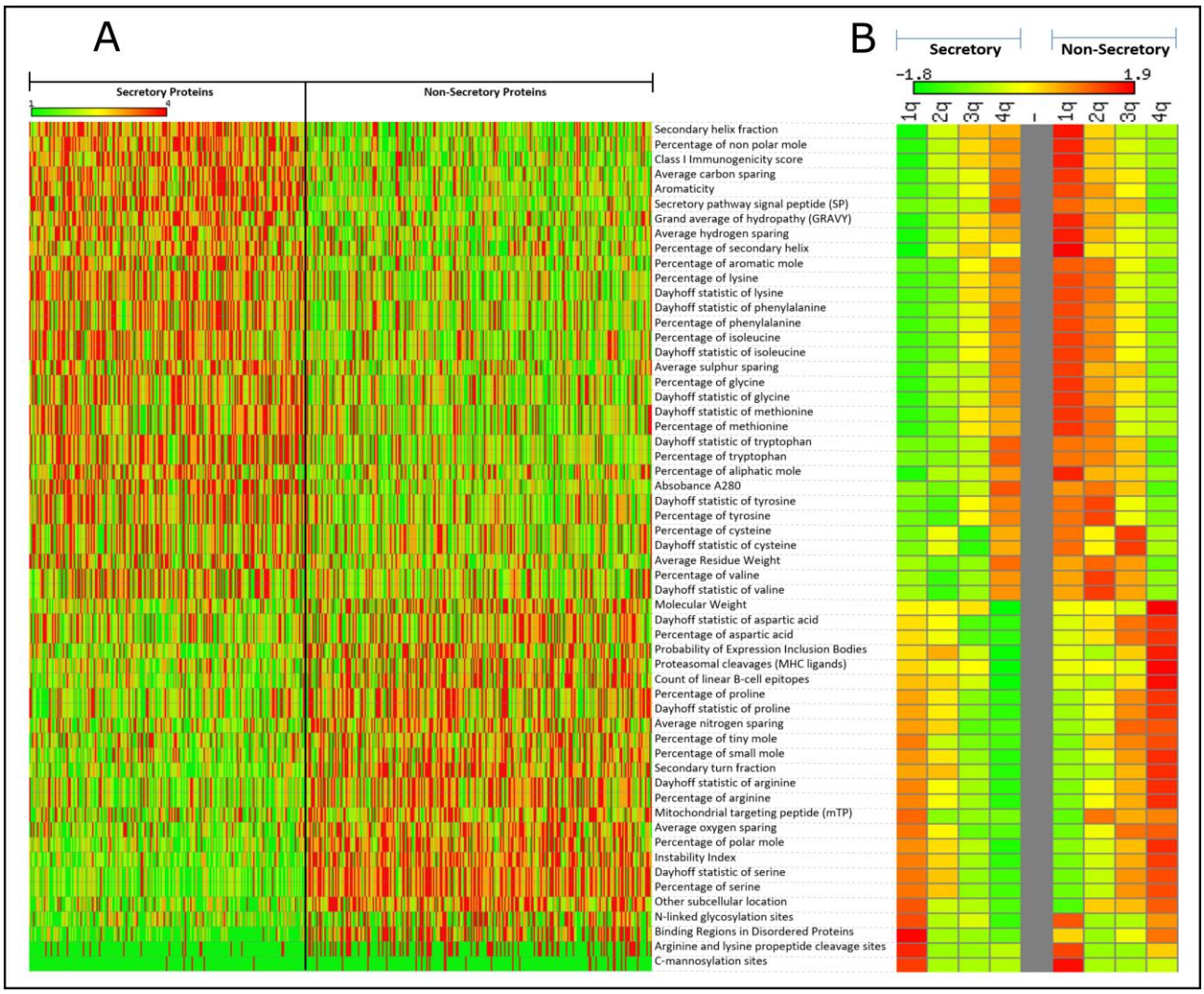


Figure 2.3 Features associated with secreted peptides.

Means in the secretory positive and secretory negative training sets were compared by t-test. Shown are all features with p<0.01. (A) Heatmap of features significantly differentially distributed between secreted and non-secreted peptides. Columns represent each protein of the training set, rows represent features. (B) Quantiles distribution of significantly different features for secreted and non-secreted peptides. Values are depicted in color code, ranging from green (low) to red (high).

Table 2.2 Features differentially distributed between surface and non-surface proteins.

Means between positive and negative training sets were compared by t-test. Shown are all features with p<0.01. P-values were corrected for multiple testing using Bonferroni correction and False Discovery Rate (FDR).

Features	Mean surface proteins	Mean non-surface proteins	P-value	Bonferroni corrected P-value	FDR

Dayhoff statistic of serine	0.9535	1.4182	6.76E-44	5.48E-42	2.77E-42
Percentage of serine	6.6747	9.9273	6.85E-44	5.55E-42	2.77E-42
Secondary turn fraction	0.2081	0.2614	1.69E-42	1.37E-40	4.57E-41
Instability Index	35.4030	47.1189	1.40E-33	1.13E-31	2.84E-32
Proteasomal cleavages (MHC ligands)	82.6185	167.0181	5.58E-30	4.52E-28	9.03E-29
Molecular Weight	29115.7535	57811.0814	5.66E-29	4.59E-27	7.65E-28
Count of linear B-cell epitopes	76.3253	187.0397	8.47E-28	6.86E-26	9.80E-27
Dayhoff statistic of proline	0.6589	0.9940	3.59E-23	2.91E-21	3.63E-22
Percentage of proline	3.4262	5.1686	2.00E-19	1.62E-17	1.80E-18
Probability of Expression Inclusion Bodies	0.6906	0.7873	5.49E-17	4.45E-15	4.45E-16
N-linked glycosylation sites	1.1566	3.5957	3.29E-16	2.67E-14	2.42E-15
Secretory pathway signal peptide (SP)	0.3167	0.1128	5.06E-16	4.10E-14	3.42E-15
Dayhoff statistic of lysine	1.1220	0.8335	1.14E-14	9.23E-13	7.10E-14
Molar Extinction Coefficient A280	29577.2691	57284.1516	1.24E-14	1.00E-12	7.17E-14
Percentage of lysine	7.4053	5.5013	1.77E-13	1.43E-11	9.54E-13
Mitochondrial targeting peptide (mTP)	0.1514	0.2827	3.26E-13	2.64E-11	1.65E-12
Binding Regions in Disordered Proteins	0.6426	3.0361	7.32E-13	5.93E-11	3.49E-12
Dayhoff statistic of asparagine	1.0742	1.3795	1.14E-12	9.19E-11	5.11E-12
Arginine and lysine propeptide cleavage sites	0.0683	0.4188	1.27E-11	1.03E-09	5.43E-11
Percentage of polar mole	46.2646	50.2015	2.59E-11	2.10E-09	1.05E-10
Dayhoff statistic of histidine	1.0947	1.4063	2.77E-11	2.24E-09	1.07E-10
Percentage of non polar mole	53.7354	49.7985	3.06E-11	2.48E-09	1.13E-10
Average hydrogen sparing	9.9920	9.8428	1.12E-10	9.09E-09	3.79E-10
Percentage of histidine	2.1894	2.8127	1.12E-10	9.10E-09	3.79E-10
Percentage of asparagine	4.6189	5.9321	1.52E-10	1.23E-08	4.84E-10

Dayhoff statistic of methionine	1.5714	1.1603	1.55E-10	1.26E-08	4.84E-10
Percentage of secondary helix	41.9157	30.2960	2.22E-10	1.80E-08	6.67E-10
Percentage of small mole	47.3355	50.2022	2.45E-10	1.98E-08	7.08E-10
Percentage of methionine	2.6714	1.9725	1.04E-09	8.40E-08	2.90E-09
Average oxygen sparing	2.4703	2.5076	1.17E-08	9.48E-07	3.16E-08
Average nitrogen sparing	1.3538	1.3879	2.96E-08	2.39E-06	7.73E-08
Percentage of arginine	4.7430	5.5201	9.81E-08	7.95E-06	2.42E-07
Dayhoff statistic of arginine	0.9680	1.1265	9.85E-08	7.98E-06	2.42E-07
Average sulphur sparing	0.0500	0.0405	1.06E-07	8.62E-06	2.54E-07
Percentage of isoleucine	7.2918	5.9486	2.28E-07	1.85E-05	5.17E-07
Dayhoff statistic of isoleucine	1.6204	1.3219	2.30E-07	1.86E-05	5.17E-07
Secondary helix fraction	0.3279	0.3015	6.44E-07	5.22E-05	1.41E-06
Percentage of aliphatic mole	23.5608	21.5818	3.17E-06	0.0003	6.76E-06
Secondary sheet fraction	0.2465	0.2278	4.79E-06	0.0004	9.95E-06
Dayhoff statistic of valine	1.0390	0.8925	5.81E-06	0.0005	1.15E-05
Percentage of valine	6.8571	5.8904	5.83E-06	0.0005	1.15E-05
Grand average of hydropathy (GRAVY)	-0.1669	-0.3786	2.46E-05	0.0020	4.75E-05
Percentage of alanine	6.4305	5.3151	0.0001	0.0087	0.0002
Dayhoff statistic of alanine	0.7477	0.6180	0.0001	0.0088	0.0002
Percentage of glycine	6.0926	5.1145	0.0004	0.0302	0.0007
Dayhoff statistic of glycine	0.7253	0.6089	0.0004	0.0304	0.0007
Average carbon sparing	5.0559	4.9950	0.0006	0.0499	0.0011
Class I Immunogenicity score	0.1592	-2.6561	0.0007	0.0530	0.0011
C-mannosylation sites	0.0281	0.1047	0.0010	0.0840	0.0017
Dayhoff statistic of glutamine	0.9200	1.0214	0.0022	0.1771	0.0035
Percentage of tryptophan	1.1583	1.1895	0.0034	0.2773	0.0054
Percentage of glutamine	3.5879	3.9833	0.0038	0.3038	0.0058
Other subcellular location	0.5970	0.6544	0.0048	0.3925	0.0074
Dayhoff statistic of tryptophan	0.8910	0.9150	0.0049	0.4001	0.0074

Table 2.3 Features differentially distributed between secretory and non-secretory proteins.

Means between positive and negative training sets were compared by t-test and p-values were adjusted for multiple testing using Bonferroni correction and False Discovery Rate (FDR).

Features	Mean secretory proteins	Mean non-secretory proteins	P-value	Bonferroni corrected P-value	FDR
Dayhoff statistic of serine	0.9223	1.5033	1.24E-37	1.00E-35	1.00E-35
Percentage of serine	6.4565	10.5232	5.45E-31	4.41E-29	2.21E-29
Binding Regions in Disordered Proteins	0.6829	6.7519	8.56E-31	6.93E-29	2.31E-29
Instability Index	37.3205	49.4272	3.61E-30	2.92E-28	7.31E-29
Secretory pathway signal peptide (SP)	0.4327	0.0876	9.31E-30	7.54E-28	1.51E-28
Percentage of non polar mole	52.4495	46.9683	6.92E-26	5.61E-24	9.35E-25
Other subcellular location	0.4646	0.7877	1.30E-24	1.05E-22	1.50E-23
Secondary helix fraction	0.3092	0.2698	4.20E-24	3.40E-22	4.25E-23
Percentage of polar mole	47.5505	53.0317	8.95E-24	7.25E-22	8.06E-23
Secondary turn fraction	0.2240	0.2662	1.13E-21	9.12E-20	9.12E-21
Aromaticity	0.0953	0.0718	7.79E-19	6.31E-17	5.74E-18
Count of linear B-cell epitopes	105.2098	285.0078	5.03E-17	4.07E-15	3.39E-16
Average carbon sparing	5.0469	4.8997	1.55E-16	1.25E-14	9.65E-16
Arginine and lysine propeptide cleavage sites	0.0976	0.6822	8.64E-14	7.00E-12	5.00E-13
Dayhoff statistic of arginine	0.9398	1.2147	3.20E-13	2.59E-11	1.64E-12
Percentage of arginine	4.6052	5.9521	3.24E-13	2.62E-11	1.64E-12
Percentage of aromatic mole	12.0057	9.9595	6.48E-13	5.25E-11	3.09E-12
Grand average of hydropathy (GRAVY)	-0.3258	-0.5314	6.24E-12	5.06E-10	2.81E-11
Percentage of small mole	48.4774	51.8143	2.22E-11	1.80E-09	9.47E-11
Average oxygen sparing	2.4879	2.5318	8.16E-11	6.61E-09	3.30E-10
Absorbance A280	1.1818	0.8254	1.84E-10	1.49E-08	7.08E-10

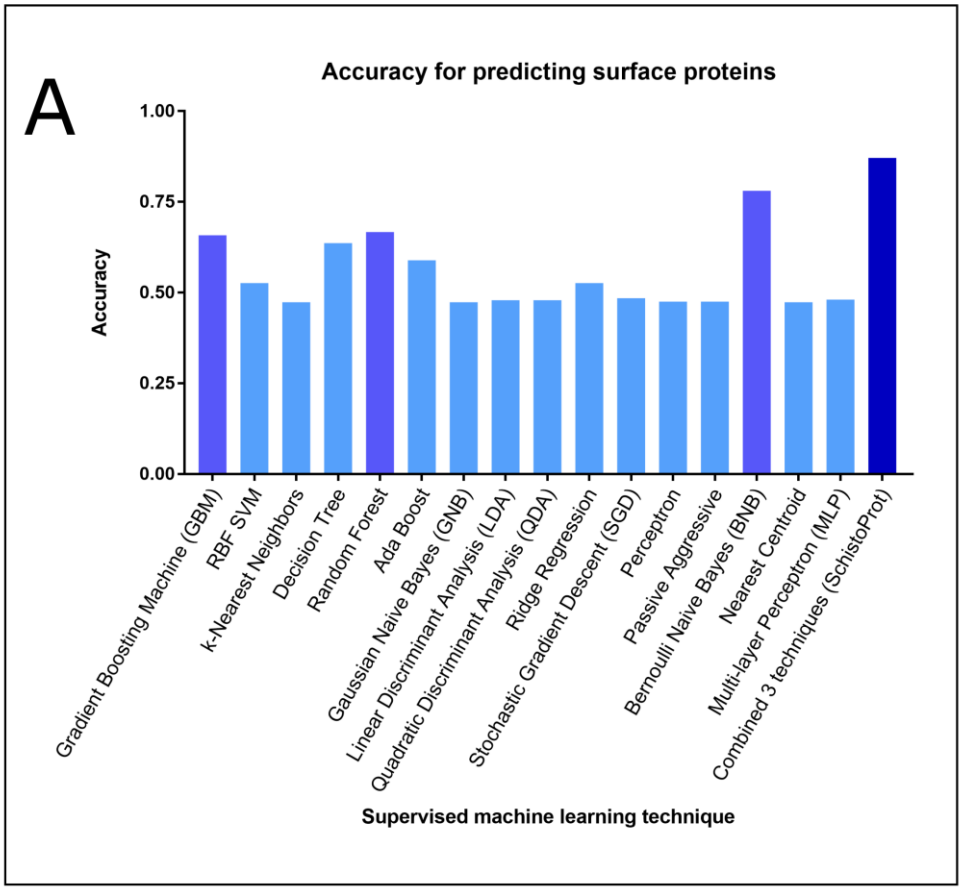
Proteasomal cleavages (MHC ligands)	102.4829	196.5853	1.23E-09	9.97E-08	4.53E-09
Dayhoff statistic of phenylalanine	1.1884	0.9220	1.86E-09	1.51E-07	6.57E-09
Probability of Expression Inclusion Bodies	0.7104	0.7896	1.96E-09	1.59E-07	6.62E-09
Percentage of phenylalanine	4.2782	3.3190	4.52E-09	3.66E-07	1.46E-08
Molecular Weight	36616.5295	70164.3000	1.21E-08	9.80E-07	3.77E-08
Mitochondrial targeting peptide (mTP)	0.1489	0.1938	1.29E-08	1.05E-06	3.88E-08
Average sulphur sparing	0.0549	0.0427	1.74E-08	1.41E-06	5.03E-08
Percentage of proline	4.1500	5.1613	2.17E-08	1.76E-06	5.89E-08
Dayhoff statistic of proline	0.7981	0.9925	2.18E-08	1.77E-06	5.89E-08
Average hydrogen sparing	9.8966	9.7668	3.51E-08	2.84E-06	9.16E-08
Average nitrogen sparing	1.3686	1.4017	7.82E-08	6.33E-06	1.98E-07
Percentage of lysine	7.2160	6.0889	8.52E-08	6.90E-06	2.03E-07
Dayhoff statistic of lysine	1.0933	0.9226	8.52E-08	6.90E-06	2.03E-07
Percentage of isoleucine	6.2476	5.4087	1.01E-07	8.17E-06	2.28E-07
Dayhoff statistic of isoleucine	1.3883	1.2019	1.01E-07	8.20E-06	2.28E-07
Dayhoff statistic of tryptophan	1.0873	0.6816	1.04E-07	8.45E-06	2.28E-07
Percentage of tryptophan	1.4135	0.8861	2.03E-07	1.65E-05	4.34E-07
Percentage of glycine	6.1741	5.0614	5.31E-07	4.30E-05	1.08E-06
Dayhoff statistic of glycine	0.7350	0.6026	5.33E-07	4.31E-05	1.08E-06
Percentage of aliphatic mole	21.3916	19.8003	8.75E-07	7.09E-05	1.73E-06
Percentage of tiny mole	27.0767	29.3396	1.00E-06	8.11E-05	1.93E-06
Class I Immunogenicity score	-0.2818	-5.8851	8.77E-06	0.00071015	1.65E-05
Dayhoff statistic of tyrosine	1.1290	0.8749	9.47E-06	0.00076723	1.74E-05
Percentage of secondary helix	37.4439	31.8217	1.42E-05	0.00114778	2.55E-05
Dayhoff statistic of methionine	1.4054	1.1315	6.38E-05	0.0051643	0.0001

Percentage of tyrosine	3.8388	2.9748	7.93E-05	0.0064232 2	0.0001
Average Residue Weight	113.3850	112.2477	9.44E-05	0.0076490 6	0.0002
Percentage of methionine	2.3891	1.9236	0.0001	0.0115401 4	0.0002
Percentage of valine	6.3702	5.6791	0.0003	0.0229345 6	0.0004
Dayhoff statistic of valine	0.9652	0.8605	0.0003	0.0229452 4	0.0004
N-linked glycosylation sites	1.8878	4.5969	0.0006	0.0518811 8	0.0010
C-mannosylation sites	0.0244	0.0930	0.0018	0.1477108 7	0.0028
Percentage of cysteine	3.0994	2.3482	0.0024	0.1953022	0.0036
Dayhoff statistic of cysteine	1.0687	0.8097	0.0029	0.2386676 5	0.0043
Dayhoff statistic of aspartic acid	0.9571	1.0481	0.0073	0.5944364	0.0105
Percentage of aspartic acid	5.2637	5.7646	0.0074	0.5966487 1	0.0105

2.5.3 Performance evaluation of 16 machine-learning techniques

The classification accuracy of 16-different supervised machine learning techniques has been evaluated. Classification performance was assessed on a training set of known *Schistosoma* surface proteins and secreted peptides using stratified k-fold (10-fold) cross-validation (Supplementary Table 2.6;

Supplementary Table 2.7). The 3 top performing methods achieved individual classification accuracies in the range of 0.65 - 0.78 for surface proteins and 0.71 - 0.80 for secreted peptides (Figure 2.4; Table 2.4). The combination of these 3 techniques achieved a superior accuracy of 87 for surface/non-surface and 94 for secretory/non-secretory classifiers (Figure 2.4; Table 2.4) and is used in the SchistoProt webserver.


A
Accuracy for predicting surface proteins
Accuracy
1.00
0.75
0.50
0.25
0.00
Gradient Boosting Machine (GBM)
RBF SVM
k-Nearest Neighbors
Decision Tree
Random Forest
Ada Boost
Gaussian Naive Bayes (GNB)
Linear Discriminant Analysis (LDA)
Quadratic Discriminant Analysis (QDA)
Ridge Regression
Stochastic Gradient Descent (SGD)
Perceptron
Passive Aggressive
Bernoulli Naive Bayes (BNB)
Nearest Centroid
Multi-layer Perceptron (MLP)
Combined 3 techniques (SchistoProt)
Supervised machine learning technique

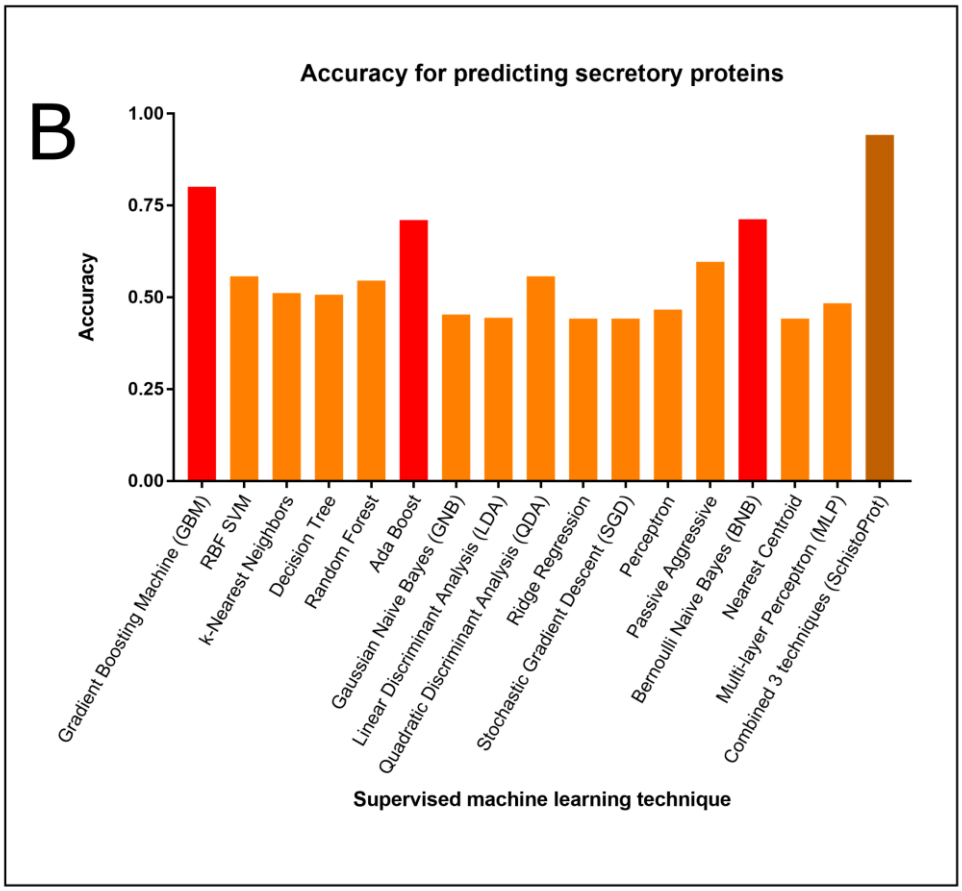

B
Accuracy for predicting secretory proteins
Accuracy
1.00
0.75
0.50
0.25
0.00
Gradient Boosting Machine (GBM)
RBF SVM
k-Nearest Neighbors
Decision Tree
Random Forest
Ada Boost
Gaussian Naive Bayes (GNB)
Linear Discriminant Analysis (LDA)
Quadratic Discriminant Analysis (QDA)
Ridge Regression
Stochastic Gradient Descent (SGD)
Perceptron
Passive Aggressive
Bernoulli Naive Bayes (BNB)
Nearest Centroid
Multi-layer Perceptron (MLP)
Combined 3 techniques (SchistoProt)
Supervised machine learning technique

Figure 2.4 Comparison of different supervised machine learning techniques for the identification of Schistosoma surface proteins and secreted peptides.

(A) Prediction accuracy for surface proteins. (B) Prediction accuracy for secreted peptides. Classifiers were trained on the training set of known *Schistosoma* surface proteins and secreted peptides and evaluated by stratified k-fold (10-fold) cross-validation.

Table 2.4 Comparison of prediction accuracy of 16 supervised machine learning techniques.

Classifiers were evaluated on the training set of known surface (n=249), non-surface (n=277) and known secreted (n=205), non-secreted (n=258) proteins by stratified k-fold (10-fold) cross-validation. Additionally, the classification accuracy of the SchistoProt classifier was evaluated, which is based on the combination of the high performing 3 machine learning techniques. AUC: Area Under the Roc Curve.

Machine Learning Technique	Surface Classification Overall Accuracy	Surface Classification AUC	Secretory Classification Overall Accuracy	Secretory Classification AUC
Gradient Boosting Machine (GBM)	0.6581	0.6474	0.8015	0.8027
RBF SVM	0.5266	0.5000	0.5573	0.5000
k-Nearest Neighbors	0.4734	0.5000	0.5118	0.5000
Decision Tree	0.6366	0.6330	0.5077	0.4923
Random Forest	0.6670	0.6583	0.5460	0.5176
Ada Boost	0.5889	0.6082	0.7107	0.7301
Gaussian Naive Bayes (GNB)	0.4734	0.5000	0.4534	0.5000
Linear Discriminant Analysis (LDA)	0.4791	0.5000	0.4449	0.5005
Quadratic Discriminant Analysis (QDA)	0.4791	0.5000	0.5573	0.5000
Ridge Regression	0.5266	0.5000	0.4427	0.5000
Stochastic Gradient Descent (SGD)	0.4848	0.5108	0.4427	0.5000
Perceptron	0.4753	0.5018	0.4671	0.5219
Passive Aggressive	0.4753	0.5018	0.5967	0.6299
Bernoulli Naive Bayes (BNB)	0.7809	0.7837	0.7128	0.7173
Nearest Centroid	0.4734	0.5000	0.4427	0.5000
Multi-layer Perceptron (MLP)	0.4809	0.5011	0.4844	0.5106
Combined 3 techniques (SchistoProt)	0.8715	0.8647	0.9425	0.9721

2.5.4 Prediction accuracy of SchistoProt evaluated on independent test set

The performance of SchistoProt was first evaluated on the training set by stratified 10-fold cross-validation (Figure 2.4; Table 4). The final classifier (trained on the entire training set)

was then evaluated on an independent test set of 400 *Schistosoma* proteins (T400). The classification accuracy for surface proteins was compared with Phobius[45] and TMHMM[47], two general hidden Markov model based tools for the identification of surface proteins. Classification accuracy for secretory peptides was compared with SignalP[48], Phobius[45] and PrediSi[49]. For surface proteins, SchistoProt achieved a sensitivity, specificity and overall accuracy of 0.85, 0.81 and 0.83, respectively (Table 2.5). For secretory proteins sensitivity, specificity and overall accuracy were 0.92, 0.93 and 0.93, respectively (Table 2.5). SchistoProt showed a significantly higher prediction accuracy compared to the existing tools Phobius, TMHMM, PrediSi and SignalP.

Table 2.5 Performance comparison with existing prediction tools.

Prediction accuracy of existing tools and SchistoProt was evaluated on the training set of 249 surface proteins, 277 non-surface proteins, 205 secreted and 258 non-secreted peptides. Additionally, the classification accuracy was evaluated on the independent test set of 100 surface proteins, 100 non-surface proteins, 100 secreted and 100 non-secreted peptides (T400).

Surface Proteins								
Tool	Dataset	True Positive	True Negative	False Positive	False Negative	Sensitivity	Specificity	Overall Accuracy
Phobius	Training set	75	236	41	174	0.30	0.85	0.59
	Test set	27	82	18	73	0.27	0.82	0.55
Philius	Training set	67	236	41	182	0.27	0.85	0.58
	Test set	23	85	15	77	0.23	0.85	0.54
TMHMM	Training set	72	244	33	177	0.29	0.88	0.60
	Test set	24	86	14	76	0.24	0.86	0.51
SchistoProt	Training set	249	277	0	0	1	1	1
	Test set	85	81	19	15	0.85	0.81	0.83
Secreted Peptides								
Tool	Dataset	True Positive	True Negative	False Positive	False Negative	Sensitivity	Specificity	Overall Accuracy
Phobius	Training set	66	255	3	139	0.32	0.99	0.69
	Test set	24	99	1	76	0.24	0.99	0.62
SignalP	Training set	61	256	2	144	0.30	0.99	0.68
	Test set	19	99	1	81	0.19	0.99	0.59
PrediSi	Training set	60	255	3	145	0.29	0.99	0.68

	Test set	24	100	0	76	0.24	1	0.62
SchistoProt	Training set	205	258	0	0	1	1	1
	Test set	92	93	7	8	0.92	0.93	0.93

2.5.5 User-interface

SchistoProt provides an easy-to-use graphical user interface (GUI), an extensive help page and user forum page. As input, multiple protein sequences can be uploaded or pasted in fasta format. By default, SchistoProt includes 183 selected features (54 of biochemical and structural features and 129 of 2-mers) for surface proteins prediction and 179 selected features (57 of biochemical and structural features and 122 of 2-mers) for secretory peptides prediction (More Conservative mode). However, optional all available 481 features can be used (Less Conservative mode). An interactive results page is generated. A table lists the sequence ID of each query sequence, the prediction (surface/non-surface or secreted/non-secreted) and classification score (number of positive classifiers). A second table lists the individual predictions obtained for each of the 3 classifiers. Additionally, the decision score and probability are shown. The distribution of sequence features in each query protein are presented in a table and in interactive charts and plots (strip chart, heatmap and bar chart) (Figure 2.5 and Figure 2.6).

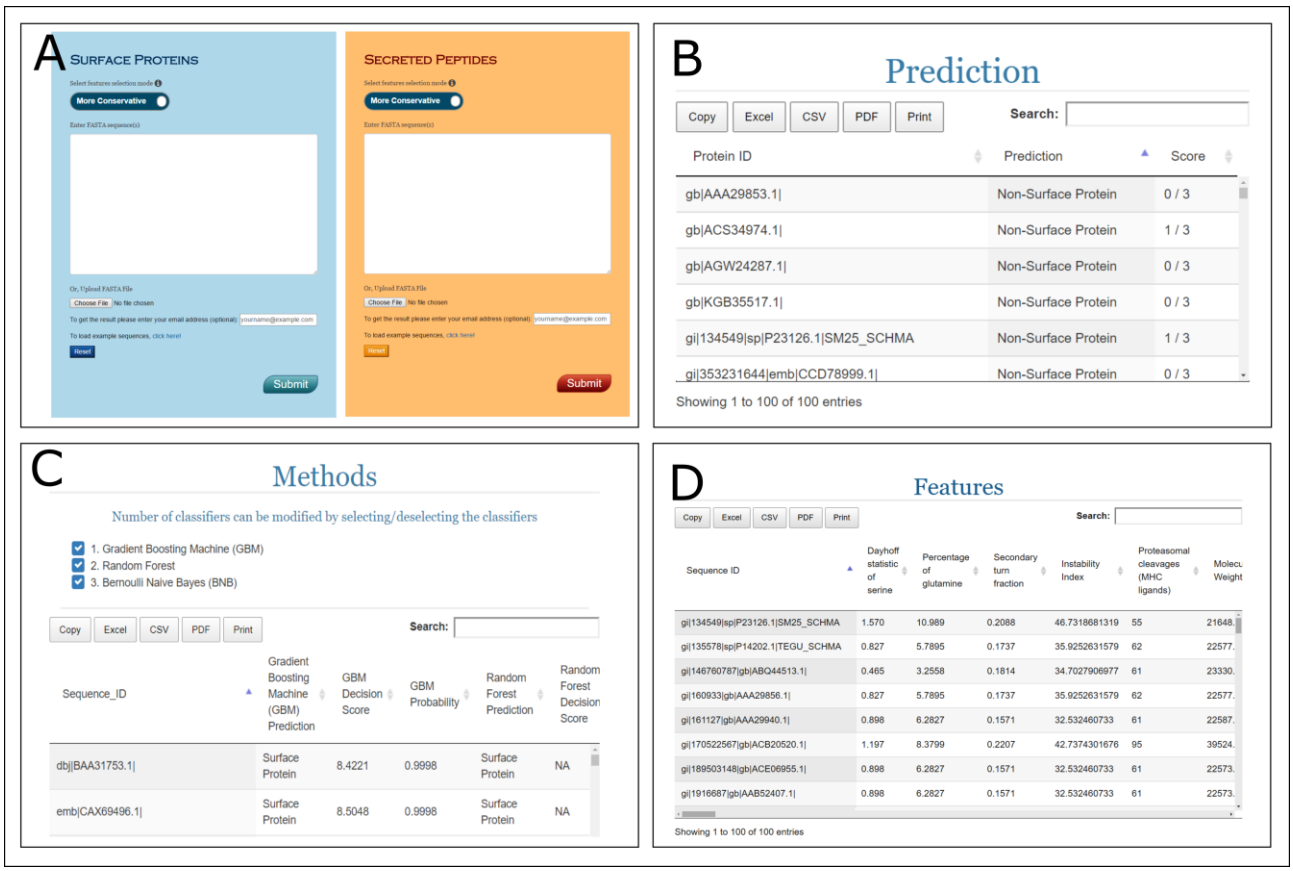


Figure 2.5 SchistoProt graphical user interface (GUI).

(A) Input forms for upload query protein sequences; (B) SchistoProt predictions are presented as table with protein ID, prediction class and score. Additionally, the decision score and probability are presented (not shown in screenshot). (C) Individual predictions obtained for each of the 3 used machine-learning techniques. Shown are the ID of each query sequence, prediction class, decision score and probability; (D) Feature table presenting the frequencies of the features in each query protein.

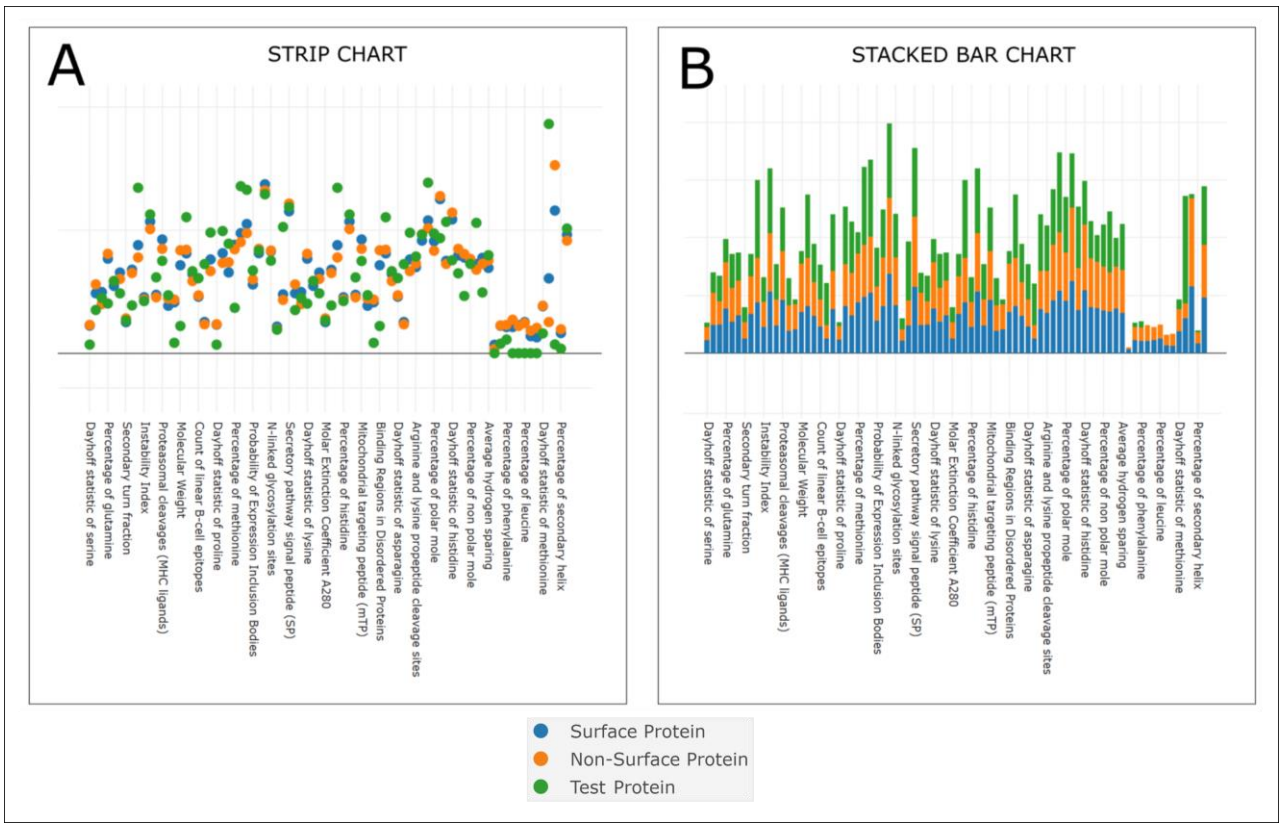


Figure 2.6 Graphical presentation of SchistoProt predictions.

SchistoProt presents results as (A) Strip Chart and (B) Stacked Bar. For each feature and each selected query protein the figures show the mean frequency of the feature in the positive and negative training sets and the frequency of the feature in the selected protein.

2.5.6 Architecture and run-time performance

The SchistoProt server is developed in Python using the Django web framework (Figure 2.7). The server can handle whole-proteome datasets and there is no limit for the number of uploaded query sequences. The server performs background task processing and can process multiple user sessions in parallel. After data submission a link is provided, which gives access to the predictions if the users browser window has been closed.

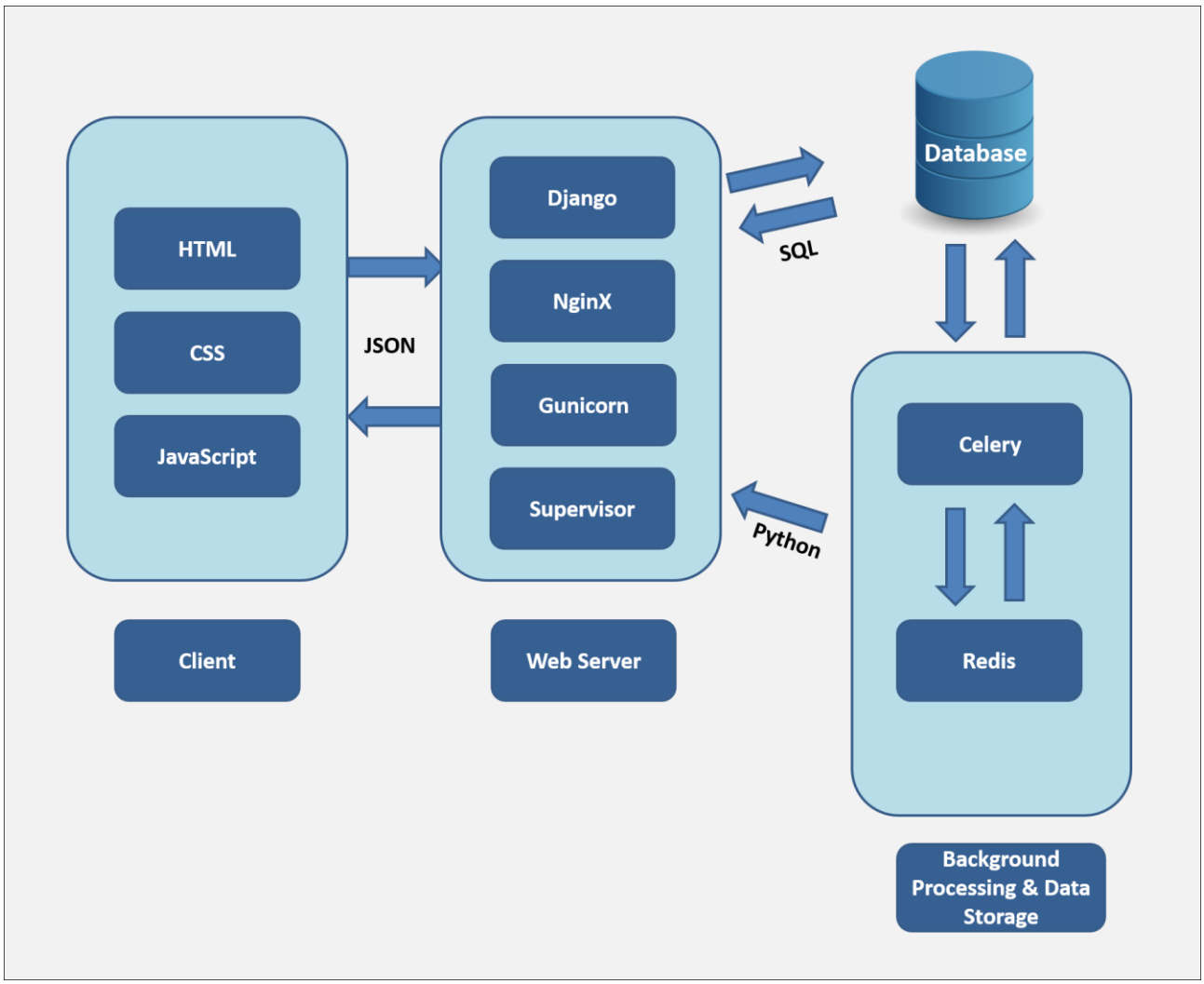


Figure 2.7 ShistoProt web server architecture.

SchistoProt has been designed following a 3 tiers framework: i) The client interface, ii) The web server and iii) Data storage and background processing. The client interface has been realized using HTML web-pages, cascading style sheets (CSS) and JavaScript. The web server relies on the Django Python web framework and the NginX high-performance HTTP-server. Gunicorn and Supervisor are used for managing and running multiple workers. Background processing has been realized via Celery and Redis. Data is persistently stored in a SQLite database.

Computed results are stored in a SQLite database for later re-use. This saves considerable runtime if the same sequences are uploaded multiple times, e.g. by different users. SchistoProt requires 227.84 seconds for processing 100 query sequences if the sequences are not found in the database, whereas it takes only 1.16 seconds if all 100 sequences are present in the database.

2.6 Conclusion

SchistoProt is an easy-to-use, accurate and fast classifier for the *in silico* identification of *Schistosoma* surface proteins and secreted peptides. The software has been optimized for large datasets and allows rapid whole-proteome analysis. The obtained results demonstrate that a genus-specific classifier is superior to general tools for the *in silico* prediction of *Schistosoma* surface proteins. Furthermore, as others have also found[27], a genus-specific classifier is also superior for the identification of secreted proteins. SchistoProt assists researchers in identifying genes important for host-parasite interaction, studying anti-schistosome protective immunity, and identifying candidate vaccine targets. It therefore represents a valuable tool for improving our understanding of *Schistosoma* pathogenicity and host-parasite interaction, and for informing the rational design of much-needed *Schistosoma* vaccines.

Supporting information

Supplementary Text 2.1 16 different supervised machine learning classifiers and their parameter settings used to select optimum classifiers in SchistoProt.

Classifiers are run using the Scikit-learn (Version 0.18.1) library in Python with default parameters.

Gradient Boosting Machine (GBM). Gradient Boosting Method (GBM) is an additive model in a forward stage-wise fashion; it allows for the optimization of arbitrary differentiable loss functions.an additive model in a forward stage-wise fashion; it allows for the optimization of arbitrary differentiable loss functions. The parameters settings were loss='deviance', learning_rate=0.1, n_estimators=100, subsample=1.0, criterion='friedman_mse', min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_depth=3, min_impurity_split=1e-07, init=None, random_state=None, max_features=None, verbose=0, max_leaf_nodes=None, warm_start=False, presort='auto'.

RBF SVM. Support Vector Machine (SVM) is a classifier which learns by finding the separating hyperplane that maximizes the margin between two classes of a training set. RBF SVM is Support Vector Machine with radial bias function (RBF) kernel. The parameters settings were C=12, cache_size=200, class_weight=None, coef0=0.0, decision_function_shape=None, degree=3, gamma=2, kernel='rbf', max_iter=-1, probability=False, random_state=None, shrinking=True, tol=0.001 and verbose=False.

k-Nearest Neighbors. The principle behind k-nearest neighbor methods is to find a predefined number of training samples closest in distance to the new point, and predict the label from these. The parameters settings were algorithm='auto', leaf_size=30, metric='minkowski', metric_params=None, n_jobs=1, n_neighbors=12, p=2 and weights='uniform'.

Decision Tree. Decision Trees are a non-parametric supervised learning method used for classification. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. The parameters settings were class_weight=None, criterion='gini', max_depth=12, max_features=None, max_leaf_nodes=None, min_samples_leaf=1, min_samples_split=2, min_weight_fraction_leaf=0.0, presort=False, random_state=None and splitter='best'.

Random Forest. A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and use averaging to improve the predictive accuracy and control over-fitting. The parameters settings were bootstrap=True, class_weight=None, criterion='gini', max_depth=12, max_features=80, max_leaf_nodes=None, min_samples_leaf=1, min_samples_split=2, min_weight_fraction_leaf=0.0, n_estimators=12, n_jobs=1, oob_score=False, random_state=None, verbose=0 and warm_start=False.

Ada Boost. An Ada Boost classifier is a meta-estimator that begins by fitting a classifier on the original dataset and then fits additional copies of the classifier on the same dataset but where the weights of incorrectly classified instances are adjusted such that subsequent classifiers focus more on difficult cases. The parameters settings were algorithm='SAMME.R', base_estimator=None, learning_rate=1.0, n_estimators=50 and random_state=None.

Naive Bayes. Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes' theorem with the "naive" assumption of independence between every pair of features. The parameters settings were the default parameters of GaussianNB() in Scikit-learn.

Linear Discriminant Analysis (LDA). Linear Discriminant Analysis (LDA) tries to identify attributes that account for the most variance between classes by a linear surface. The parameters settings were n_components=None, priors=None, shrinkage=None, solver='svd', store_covariance=False and tol=0.0001.

Quadratic Discriminant Analysis (QDA). Quadratic Discriminant Analysis (QDA) is used in machine learning classification to separate measurements of two or more classes of objects or events by a quadric surface. The parameters settings were priors=None, reg_param=0.0, store_covariances=False and tol=0.0001.

Ridge Regression. Ridge Regression is a technique for analyzing multiple regression data that suffer from multicollinearity. The parameters settings were alpha=1.0, class_weight=None, copy_X=True, fit_intercept=True, max_iter=None, normalize=False, random_state=None, solver='auto' and tol=0.001.

Stochastic Gradient Descent (SGD). Stochastic Gradient Descent (SGD) is a simple yet very efficient approach to discriminative learning of linear classifiers under convex loss functions such as (linear) Support Vector Machines and Logistic Regression. The parameters settings were alpha=0.0001, average=False, class_weight=None, epsilon=0.1, eta0=0.0, fit_intercept=True, l1_ratio=0.15, learning_rate='optimal', loss='hinge', n_iter=5, n_jobs=1, penalty='l2', power_t=0.5, random_state=None, shuffle=True, verbose=0 and warm_start=False.

Perceptron. Perceptron is a type of linear classifier that makes its predictions based on a linear predictor function combining a set of weights with the feature vector. It does not require a learning rate and it is not regularized (penalized). The parameters settings were alpha=0.0001, class_weight=None, eta0=1.0, fit_intercept=True, n_iter=5, n_jobs=1, penalty=None, random_state=0, shuffle=True, verbose=0 and warm_start=False.

Passive Aggressive. The passive aggressive algorithms are a family of algorithms for large-scale learning. They are similar to the Perceptron in that they do not require a learning rate. However, contrary to the Perceptron, they include a regularization parameter C. The parameters settings were C=1.0, class_weight=None, fit_intercept=True,

loss='hinge', n_iter=5, n_jobs=1, random_state=None, shuffle=True, verbose=0 and warm_start=False.

Bernoulli Naive Bayes. Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes' theorem with the "naive" assumption of independence between every pair of features. Bernoulli Naive Bayes implements the naive Bayes training and classification algorithms for data that is distributed according to multivariate Bernoulli distributions; i.e., there may be multiple features but each one is assumed to be a binary-valued (Bernoulli, boolean) variable. Therefore, this class requires samples to be represented as binary-valued feature vectors; if handed any other kind of data, a Bernoulli Naive Bayes instance may binarize its input (depending on the binarize parameter). The parameters settings were alpha=1.0, binarize=0.0, class_prior=None and fit_prior=True.

Nearest Centroid. Nearest Centroid classifier is a classification model that assigns to observations the label of the class of training samples whose mean (centroid) is closest to the observation. The parameters settings were metric='euclidean' and shrink_threshold=None.

Multi-layer Perceptron (MLP). A multilayer perceptron (MLP) is a network of simple neurons called perceptrons. The perceptron computes a single output from multiple real-valued inputs by forming a linear combination according to its input weights and then possibly putting the output through some nonlinear activation function. The parameters settings were hidden_layer_sizes=(100,), activation='relu', solver='adam', alpha=0.0001, batch_size='auto', learning_rate='constant', learning_rate_init=0.001, power_t=0.5, max_iter=200, shuffle=True, random_state=None, tol=0.0001, verbose=False, warm_start=False, momentum=0.9, nesterovs_momentum=True, early_stopping=False, validation_fraction=0.1, beta_1=0.9, beta_2=0.999, epsilon=1e-08.

Supplementary Table 2.1 List of 81 features used in SchistoProt for protein classification. SchistoProt uses 481 features for protein classification.

Of these 481 features, 81 features represent biochemical and structural properties (shown in this table). The remaining 400 features represent 2-mers of the 20 amino acids (Supplementary Table 2.2).

Percentage of alanine	Secondary sheet fraction	DayhoffStat of threonine
Percentage of cysteine	Average Residue Weight	DayhoffStat of valine
Percentage of aspartic acid	Average carbon sparing	DayhoffStat of tryptophan
Percentage of glutamic acid	Average nitrogen sparing	DayhoffStat of tyrosine
Percentage of phenylalanine	Average sulphur sparing	Percentage of tiny mole
Percentage of glycine	Average oxygen sparing	Percentage of small mole
Percentage of histidine	Average hydrogen sparing	Percentage of aliphatic mole
Percentage of isoleucine	Charge	Percentage of aromatic mole
Percentage of lysine	Molar Extinction Coefficient A280	Percentage of polar mole
Percentage of leucine	Absobance A280	Percentage of non polar mole
Percentage of methionine	Probability of Expression Inclusion Bodies	Percentage of charged mole
Percentage of asparagine	DayhoffStat of alanine	Percentage of acidic mole
Percentage of proline	DayhoffStat of cysteine	Percentage of basic mole
Percentage of glutamine	DayhoffStat of aspartic acid	Percentage of secondary helix
Percentage of arginine	DayhoffStat of glutamic acid	Percentage of secondary sheet
Percentage of serine	DayhoffStat of phenylalanine	Percentage of secondary turns
Percentage of threonine	DayhoffStat of glycine	Percentage of secondary coil
Percentage of valine	DayhoffStat of histidine	C-mannosylation sites
Percentage of tryptophan	DayhoffStat of isoleucine	Proteasomal cleavages (MHC ligands)
Percentage of tyrosine	DayhoffStat of lysine	N-linked glycosylation sites
Molecular Weight	DayhoffStat of leucine	Arginine and lysine propeptide cleavage sites
Aromaticity	DayhoffStat of methionine	Binding Regions in Disordered Proteins

Instability Index	DayhoffStat of asparagine	Mitochondrial targeting peptide (mTP)
Isoelectric Point	DayhoffStat of proline	Secretory pathway signal peptide (SP)
Grand average of hydropathy (GRAVY)	DayhoffStat of glutamine	Other subcellular location
Secondary helix fraction	DayhoffStat of arginine	Linear B-cell epitopes
Secondary turn fraction	DayhoffStat of serine	Class I Immunogenicity Score

Supplementary Table 2.2 List of 400 2-mers used in SchistoProt for protein classification.

SchistoProt uses 400 2-mers of the 20 amino acids for protein classification.

AA	DA	FA	HA	KA	MA	PA	RA	TA	WA
AC	DC	FC	HC	KC	MC	PC	RC	TC	WC
AD	DD	FD	HD	KD	MD	PD	RD	TD	WD
AE	DE	FE	HE	KE	ME	PE	RE	TE	WE
AF	DF	FF	HF	KF	MF	PF	RF	TF	WF
AG	DG	FG	HG	KG	MG	PG	RG	TG	WG
AH	DH	FH	HH	KH	MH	PH	RH	TH	WH
AI	DI	FI	HI	KI	MI	PI	RI	TI	WI
AK	DK	FK	HK	KK	MK	PK	RK	TK	WK
AL	DL	FL	HL	KL	ML	PL	RL	TL	WL
AM	DM	FM	HM	KM	MM	PM	RM	TM	WM
AN	DN	FN	HN	KN	MN	PN	RN	TN	WN
AP	DP	FP	HP	KP	MP	PP	RP	TP	WP
AQ	DQ	FQ	HQ	KQ	MQ	PQ	RQ	TQ	WQ
AR	DR	FR	HR	KR	MR	PR	RR	TR	WR
AS	DS	FS	HS	KS	MS	PS	RS	TS	WS
AT	DT	FT	HT	KT	MT	PT	RT	TT	WT
AV	DV	FV	HV	KV	MV	PV	RV	TV	WV
AW	DW	FW	HW	KW	MW	PW	RW	TW	WW
AY	DY	FY	HY	KY	MY	PY	RY	TY	WY

CA	EA	GA	IA	LA	NA	QA	SA	VA	YA
CC	EC	GC	IC	LC	NC	QC	SC	VC	YC
CD	ED	GD	ID	LD	ND	QD	SD	VD	YD
CE	EE	GE	IE	LE	NE	QE	SE	VE	YE
CF	EF	GF	IF	LF	NF	QF	SF	VF	YF
CG	EG	GG	IG	LG	NG	QG	SG	VG	YG
CH	EH	GH	IH	LH	NH	QH	SH	VH	YH
CI	EI	GI	II	LI	NI	QI	SI	VI	YI
CK	EK	GK	IK	LK	NK	QK	SK	VK	YK
CL	EL	GL	IL	LL	NL	QL	SL	VL	YL
CM	EM	GM	IM	LM	NM	QM	SM	VM	YM
CN	EN	GN	IN	LN	NN	QN	SN	VN	YN
CP	EP	GP	IP	LP	NP	QP	SP	VP	YP
CQ	EQ	GQ	IQ	LQ	NQ	QQ	SQ	VQ	YQ
CR	ER	GR	IR	LR	NR	QR	SR	VR	YR
CS	ES	GS	IS	LS	NS	QS	SS	VS	YS
CT	ET	GT	IT	LT	NT	QT	ST	VT	YT
CV	EV	GV	IV	LV	NV	QV	SV	VV	YV
CW	EW	GW	IW	LW	NW	QW	SW	VW	YW
CY	EY	GY	IY	LY	NY	QY	SY	VY	YY

Supplementary Table 2.3 Test for normality of extracted features.

Extracted features were reasonably normally distributed and evaluated by mean, median and shape of the data.

	Surface Positive		Surface Negative		Secretory Positive		Surface Negative	
Features	Mean	Median	Mean	Median	Mean	Median	Mean	Median
Percentage of alanine	6.4305	6.3830	5.3151	5.2811	5.7148	5.3055	5.494	5.235
Percentage of cysteine	2.3275	1.9048	2.0747	2.0356	3.0994	2.1739	2.348	2.141
Percentage of aspartic acid	5.3793	5.4968	5.3024	5.2910	5.2637	5.2910	5.765	5.788
Percentage of glutamic acid	6.1375	6.0150	5.7452	5.5215	6.4329	6.0748	6.226	5.649
Percentage of	4.4845	4.2254	4.0333	3.8095	4.2782	4.0426	3.319	3.103

phenylalanine								
Percentage of glycine	6.0926	5.9211	5.1145	4.7619	6.1741	5.7971	5.061	4.594
Percentage of histidine	2.1894	1.9737	2.8127	2.7907	2.4751	2.3936	2.780	2.853
Percentage of isoleucine	7.2918	7.0270	5.9486	5.8824	6.2476	6.2500	5.409	5.323
Percentage of lysine	7.4053	7.1560	5.5013	5.3299	7.2160	7.1713	6.089	5.649
Percentage of leucine	9.4119	9.3750	9.7429	9.6741	8.7737	8.8106	8.713	8.751
Percentage of methionine	2.6714	2.5210	1.9725	1.8433	2.3891	2.3256	1.924	1.754
Percentage of asparagine	4.6189	4.4728	5.9321	5.6561	5.6168	5.2632	5.870	5.651
Percentage of proline	3.4262	3.3210	5.1686	4.8507	4.1500	4.1026	5.161	5.039
Percentage of glutamine	3.5879	3.3613	3.9833	3.8462	3.8524	3.7634	3.914	3.650
Percentage of arginine	4.7430	4.5249	5.5201	5.3719	4.6052	4.5564	5.952	5.718
Percentage of serine	6.6747	6.5476	9.9273	9.6927	6.4565	6.1404	10.523	10.598
Percentage of threonine	5.5285	5.3903	5.4771	5.3456	5.6319	5.4054	5.913	5.944
Percentage of valine	6.8571	7.1038	5.8904	5.8932	6.3702	5.9908	5.679	5.601
Percentage of tryptophan	1.1583	1.0204	1.1895	1.0309	1.4135	1.3575	0.886	0.826
Percentage of tyrosine	3.5834	3.5714	3.3484	3.3175	3.8388	3.7415	2.975	2.797
Molecular Weight	29115.7535	23245.0700	57811.0814	46478.2100	36616.5295	28121.0500	70164.300	39179.140
Aromaticity	0.0923	0.0946	0.0857	0.0859	0.0953	0.0961	0.072	0.069
Instability Index	35.4030	35.4633	47.1189	47.1017	37.3205	36.8936	49.427	50.182
Isoelectric Point	7.3051	7.0162	7.3483	7.1319	7.1757	7.0069	7.280	7.214
Grand average of hydropathy (GRAVY)	-0.1669	-0.2414	-0.3786	-0.3954	-0.3258	-0.3435	-0.531	-0.504
Secondary helix fraction	0.3279	0.3146	0.3015	0.3077	0.3092	0.3077	0.270	0.267
Secondary turn fraction	0.2081	0.2069	0.2614	0.2580	0.2240	0.2234	0.266	0.267
Secondary sheet fraction	0.2465	0.2444	0.2278	0.2283	0.2331	0.2296	0.224	0.222
Average Residue Weight	112.9585	112.7301	112.8086	113.1947	113.3850	113.9302	112.248	112.230
Average carbon sparing	5.0502	5.0287	4.9950	5.0138	5.0465	5.0537	4.893	4.895
Average nitrogen sparing	1.3523	1.3500	1.3879	1.3846	1.3685	1.3630	1.400	1.404
Average sulphur sparing	0.0500	0.0469	0.0405	0.0387	0.0549	0.0484	0.043	0.040
Average oxygen sparing	2.4674	2.4684	2.5076	2.5086	2.4877	2.4820	2.528	2.535
Average hydrogen sparing	9.9805	9.9663	9.8428	9.8665	9.8956	9.9000	9.753	9.748
Charge	2.8715	2.5000	5.6444	6.0000	2.6122	3.0000	6.990	4.000
Molar Extinction Coefficient A280	29577.2691	22920.0000	57284.1516	45840.0000	42351.7073	35870.0000	58138.101	33635.000
Absobance A280	1.0292	0.9830	1.0176	0.9710	1.1818	1.1370	0.825	0.831
Probability of Expression Inclusion Bodies	0.6906	0.6820	0.7873	0.8100	0.7104	0.7160	0.790	0.812
DayhoffStat of alanine	0.7477	0.7420	0.6180	0.6140	0.6645	0.6170	0.639	0.609
DayhoffStat of cysteine	0.8026	0.6570	0.7154	0.7020	1.0687	0.7500	0.810	0.739
DayhoffStat of aspartic acid	0.9780	0.9990	0.9641	0.9620	0.9571	0.9620	1.048	1.053

DayhoffStat of glutamic acid	1.0229	1.0030	0.9575	0.9200	1.0721	1.0120	1.038	0.942
DayhoffStat of phenylalanine	1.2457	1.1740	1.1204	1.0580	1.1884	1.1230	0.922	0.862
DayhoffStat of glycine	0.7253	0.7050	0.6089	0.5670	0.7350	0.6900	0.603	0.547
DayhoffStat of histidine	1.0947	0.9870	1.4063	1.3950	1.2376	1.1970	1.390	1.427
DayhoffStat of isoleucine	1.6204	1.5620	1.3219	1.3070	1.3883	1.3890	1.202	1.183
DayhoffStat of lysine	1.1220	1.0840	0.8335	0.8080	1.0933	1.0870	0.923	0.856
DayhoffStat of leucine	1.2719	1.2670	1.3166	1.3070	1.1857	1.1910	1.177	1.183
DayhoffStat of methionine	1.5714	1.4830	1.1603	1.0840	1.4054	1.3680	1.132	1.032
DayhoffStat of asparagine	1.0742	1.0400	1.3795	1.3150	1.3062	1.2240	1.365	1.314
DayhoffStat of proline	0.6589	0.6390	0.9940	0.9330	0.7981	0.7890	0.993	0.969
DayhoffStat of glutamine	0.9200	0.8620	1.0214	0.9860	0.9878	0.9650	1.004	0.936
DayhoffStat of arginine	0.9680	0.9230	1.1265	1.0960	0.9398	0.9300	1.215	1.167
DayhoffStat of serine	0.9535	0.9350	1.4182	1.3850	0.9223	0.8770	1.503	1.514
DayhoffStat of threonine	0.9063	0.8840	0.8979	0.8760	0.9233	0.8860	0.969	0.975
DayhoffStat of valine	1.0390	1.0760	0.8925	0.8930	0.9652	0.9080	0.860	0.849
DayhoffStat of tryptophan	0.8910	0.7850	0.9150	0.7930	1.0873	1.0440	0.682	0.636
DayhoffStat of tyrosine	1.0539	1.0500	0.9848	0.9760	1.1290	1.1000	0.875	0.823
Percentage of tiny mole	27.0538	27.4850	27.9087	27.5860	27.0767	26.9770	29.340	28.991
Percentage of small mole	47.3354	47.9670	50.2023	49.8570	48.4774	48.0920	51.814	51.828
Percentage of aliphatic mole	23.5608	23.2290	21.5818	21.7070	21.3916	21.4180	19.800	20.223
Percentage of aromatic mole	11.4157	11.3640	11.3840	11.1940	12.0057	12.0620	9.959	9.770
Percentage of polar mole	46.2646	46.9270	50.2014	49.9660	47.5505	47.4890	53.032	53.403
Percentage of non polar mole	53.7354	53.0730	49.7986	50.0340	52.4495	52.5110	46.968	46.598
Percentage of charged mole	25.8545	26.3160	24.8817	25.0860	25.9929	25.6100	26.811	26.244
Percentage of acidic mole	11.5168	11.6880	11.0476	11.0220	11.6966	11.4940	11.991	11.558
Percentage of basic mole	14.3378	14.2080	13.8341	13.6530	14.2963	14.2860	14.821	14.224
Percentage of secondary helix	42.3711	41.6000	30.7578	30.5000	37.8907	35.9000	32.271	29.600
Percentage of secondary sheet	27.9847	26.2000	26.3910	25.9000	24.7020	24.3000	24.460	24.900
Percentage of secondary turns	21.7229	20.7000	25.4444	25.2000	27.2054	23.9000	26.657	26.450
Percentage of secondary coil	17.0791	17.1000	22.5426	21.7000	17.6337	18.0000	22.703	22.400

C-mannosylation sites	0.0281	0.0000	0.1047	0.0000	0.0244	0.0000	0.093	0.000
Proteasomal cleavages (MHC ligands)	82.6185	66.0000	167.0181	130.0000	102.4829	81.0000	196.585	110.500
N-linked glycosylation sites	1.1566	1.0000	3.5957	2.0000	1.8878	2.0000	4.609	2.000
Arginine and lysine propeptide cleavage sites	0.0683	0.0000	0.4188	0.0000	0.0976	0.0000	0.682	0.000
Binding Regions in Disordered Proteins	0.6426	0.0000	3.0361	1.0000	0.6829	0.0000	6.752	3.000
Mitochondrial targeting peptide (mTP)	0.1514	0.0880	0.2827	0.1560	0.1489	0.0740	0.194	0.116
Secretory pathway signal peptide (SP)	0.3167	0.1060	0.1128	0.0650	0.4327	0.1650	0.088	0.065
Other subcellular location	0.5970	0.7610	0.6544	0.7660	0.4646	0.4770	0.788	0.866
Linear B-cell epitopes	76.3253	58.0000	187.0397	136.0000	105.2098	82.0000	285.008	149.000
Class I Immunogenicity Score	0.1592	0.1060	-2.6561	-1.6475	-0.2818	-0.1848	-5.885	-2.556

Supplementary Table 2.4 List of 129 2-mers differentially distributed between surface and non-surface proteins.

Means between positive and negative training sets were compared by t-test. Shown are all features with p<0.01. P-values were corrected for multiple testing using False Discovery Rate (FDR).

Features	Mean surface proteins	Mean non-surface proteins	P-value	FDR
SS	0.0048	0.0122	2.94E-28	1.18E-25
SN	0.0027	0.0060	2.23E-20	4.45E-18
PS	0.0022	0.0052	2.69E-19	3.58E-17
NS	0.0027	0.0058	3.32E-16	3.32E-14
LS	0.0062	0.0102	1.88E-14	1.49E-12
PL	0.0023	0.0048	2.23E-14	1.49E-12
SQ	0.0019	0.0040	1.11E-13	6.34E-12
SR	0.0028	0.0053	3.78E-12	1.82E-10
SP	0.0021	0.0043	4.10E-12	1.82E-10
ST	0.0037	0.0066	1.23E-11	4.92E-10
QP	0.0008	0.0023	3.34E-11	1.22E-09
LN	0.0037	0.0061	1.91E-10	6.35E-09
DS	0.0030	0.0052	3.46E-10	1.06E-08
HS	0.0015	0.0031	5.14E-10	1.47E-08

KV	0.0054	0.0027	5.86E-10	1.56E-08
GK	0.0051	0.0025	1.22E-09	3.05E-08
RR	0.0023	0.0043	1.62E-09	3.82E-08
PR	0.0011	0.0026	4.14E-09	9.19E-08
PP	0.0016	0.0039	6.93E-09	1.46E-07
SD	0.0033	0.0053	9.47E-09	1.89E-07
QS	0.0023	0.0040	2.80E-08	5.34E-07
IV	0.0052	0.0029	3.36E-08	6.11E-07
RS	0.0027	0.0046	3.78E-08	6.54E-07
II	0.0065	0.0033	4.06E-08	6.54E-07
KA	0.0055	0.0031	4.09E-08	6.54E-07
NN	0.0024	0.0045	1.06E-07	1.63E-06
GA	0.0045	0.0023	1.53E-07	2.27E-06
SE	0.0036	0.0055	2.71E-07	3.87E-06
EK	0.0054	0.0032	3.60E-07	4.97E-06
ME	0.0022	0.0009	4.51E-07	6.01E-06
LP	0.0034	0.0053	5.47E-07	7.06E-06
AM	0.0017	0.0007	1.11E-06	1.39E-05
SL	0.0066	0.0091	2.84E-06	3.43E-05
MK	0.0024	0.0012	2.92E-06	3.43E-05
VI	0.0051	0.0033	3.46E-06	3.79E-05
NP	0.0018	0.0030	3.48E-06	3.79E-05
DK	0.0046	0.0027	3.51E-06	3.79E-05
PG	0.0016	0.0028	4.49E-06	4.73E-05
KK	0.0059	0.0032	4.91E-06	5.04E-05
IA	0.0047	0.0031	5.49E-06	5.49E-05
AK	0.0045	0.0027	5.96E-06	5.82E-05
AI	0.0049	0.0031	6.79E-06	6.47E-05
SV	0.0045	0.0064	7.12E-06	6.63E-05
VP	0.0021	0.0037	8.33E-06	7.57E-05
KL	0.0076	0.0053	1.03E-05	9.16E-05
KT	0.0044	0.0027	2.60E-05	0.0002
SA	0.0035	0.0050	2.77E-05	0.0002
LG	0.0060	0.0042	3.86E-05	0.0003
PI	0.0020	0.0033	4.23E-05	0.0003
RP	0.0015	0.0025	4.33E-05	0.0003
EN	0.0026	0.0039	5.22E-05	0.0004
HL	0.0020	0.0030	5.31E-05	0.0004
PV	0.0023	0.0037	6.51E-05	0.0005

YH	0.0006	0.0013	6.55E-05	0.0005
IK	0.0053	0.0037	7.09E-05	0.0005
RL	0.0048	0.0064	7.75E-05	0.0006
YI	0.0029	0.0016	8.82E-05	0.0006
VV	0.0057	0.0039	9.21E-05	0.0006
WC	0.0000	0.0003	9.29E-05	0.0006
NH	0.0009	0.0016	9.57E-05	0.0006
TF	0.0030	0.0018	0.0001	0.0007
HN	0.0008	0.0016	0.0001	0.0007
IE	0.0045	0.0031	0.0001	0.0009
IG	0.0046	0.0030	0.0001	0.0009
AS	0.0036	0.0049	0.0002	0.0010
GI	0.0047	0.0032	0.0002	0.0012
PE	0.0021	0.0031	0.0002	0.0012
AV	0.0050	0.0034	0.0002	0.0012
VK	0.0048	0.0033	0.0002	0.0012
FI	0.0044	0.0028	0.0002	0.0013
LR	0.0042	0.0055	0.0003	0.0015
YM	0.0011	0.0005	0.0004	0.0020
TR	0.0020	0.0030	0.0005	0.0025
GD	0.0037	0.0025	0.0005	0.0025
KN	0.0040	0.0028	0.0005	0.0026
NR	0.0020	0.0028	0.0005	0.0028
TS	0.0041	0.0055	0.0005	0.0028
IF	0.0031	0.0020	0.0007	0.0036
AD	0.0032	0.0022	0.0007	0.0037
FP	0.0012	0.0022	0.0008	0.0038
YP	0.0011	0.0019	0.0010	0.0048
IY	0.0027	0.0017	0.0010	0.0049
RF	0.0015	0.0023	0.0011	0.0051
VT	0.0045	0.0032	0.0012	0.0057
MA	0.0021	0.0013	0.0013	0.0061
CC	0.0012	0.0006	0.0013	0.0061
NL	0.0043	0.0054	0.0013	0.0062
DF	0.0018	0.0026	0.0014	0.0062
IP	0.0025	0.0035	0.0014	0.0062
KF	0.0029	0.0020	0.0014	0.0064
PH	0.0007	0.0013	0.0015	0.0065
DH	0.0008	0.0014	0.0015	0.0067

HY	0.0007	0.0011	0.0016	0.0068
CV	0.0018	0.0010	0.0017	0.0071
RH	0.0011	0.0018	0.0017	0.0072
DM	0.0017	0.0010	0.0020	0.0083
DG	0.0038	0.0026	0.0026	0.0105
HQ	0.0007	0.0011	0.0027	0.0111
GH	0.0011	0.0017	0.0027	0.0111
PN	0.0020	0.0028	0.0032	0.0126
PQ	0.0013	0.0020	0.0032	0.0127
EP	0.0018	0.0026	0.0033	0.0131
FH	0.0007	0.0012	0.0034	0.0131
KG	0.0030	0.0021	0.0035	0.0133
HF	0.0009	0.0015	0.0036	0.0138
FG	0.0031	0.0022	0.0037	0.0140
WR	0.0010	0.0006	0.0038	0.0141
GV	0.0042	0.0031	0.0038	0.0141
KD	0.0039	0.0029	0.0043	0.0157
CP	0.0008	0.0013	0.0044	0.0160
IS	0.0051	0.0061	0.0045	0.0163
WS	0.0006	0.0010	0.0046	0.0163
FS	0.0031	0.0040	0.0050	0.0178
CK	0.0017	0.0010	0.0051	0.0179
DA	0.0033	0.0025	0.0063	0.0217
KC	0.0019	0.0011	0.0063	0.0217
MG	0.0015	0.0009	0.0072	0.0245
MQ	0.0011	0.0006	0.0072	0.0245
KM	0.0015	0.0010	0.0074	0.0249
QA	0.0028	0.0020	0.0075	0.0251
MD	0.0016	0.0010	0.0076	0.0252
AF	0.0032	0.0023	0.0083	0.0271
TI	0.0041	0.0031	0.0084	0.0272
NE	0.0032	0.0040	0.0085	0.0273
EM	0.0014	0.0009	0.0088	0.0281
SH	0.0016	0.0022	0.0089	0.0281
FL	0.0051	0.0039	0.0090	0.0281
VG	0.0042	0.0032	0.0090	0.0281
LV	0.0065	0.0051	0.0091	0.0284

Supplementary Table 2.5 List of 122 2-mers differentially distributed between secretory and non-secretory proteins.

Means between positive and negative training sets were compared by t-test. Shown are all features with p<0.01. P-values were corrected for multiple testing using False Discovery Rate (FDR).

Features	Mean secretory proteins	Mean non-secretory proteins	P-value	FDR
SS	0.0048	0.0151	1.19E-31	4.75E-29
TS	0.0031	0.0071	4.04E-18	8.08E-16
PS	0.0026	0.0057	3.16E-17	3.32E-15
SN	0.0031	0.0068	3.32E-17	3.32E-15
SR	0.0026	0.0056	1.56E-15	1.25E-13
ST	0.0041	0.0077	3.75E-13	2.50E-11
RR	0.0023	0.0052	8.57E-13	4.90E-11
SP	0.0026	0.0051	1.30E-11	6.50E-10
RS	0.0028	0.0053	6.73E-11	2.99E-09
HS	0.0015	0.0033	2.75E-10	1.10E-08
SA	0.0028	0.0052	3.49E-10	1.27E-08
DS	0.0031	0.0055	5.41E-10	1.80E-08
SD	0.0035	0.0058	2.32E-09	7.15E-08
SQ	0.0020	0.0038	8.95E-09	2.56E-07
RL	0.0038	0.0062	1.18E-08	3.15E-07
GQ	0.0036	0.0015	1.76E-08	4.40E-07
IS	0.0037	0.0058	2.34E-08	5.51E-07
LS	0.0065	0.0096	2.50E-08	5.56E-07
NS	0.0036	0.0062	3.31E-08	6.96E-07
NN	0.0026	0.0053	3.90E-08	7.79E-07
YG	0.0035	0.0015	1.15E-07	2.20E-06
TP	0.0021	0.0036	1.30E-07	2.36E-06
SF	0.0020	0.0035	2.78E-07	4.84E-06
IG	0.0044	0.0025	6.69E-07	1.11E-05
CG	0.0028	0.0012	7.29E-07	1.17E-05
SL	0.0058	0.0082	1.71E-06	2.64E-05
PV	0.0022	0.0037	1.87E-06	2.77E-05
IK	0.0046	0.0028	2.32E-06	3.32E-05
KN	0.0045	0.0026	2.48E-06	3.43E-05
VF	0.0030	0.0015	3.02E-06	4.00E-05
ES	0.0034	0.0051	3.10E-06	4.00E-05

HT	0.0009	0.0018	3.80E-06	4.75E-05
KM	0.0015	0.0007	4.21E-06	5.10E-05
PR	0.0017	0.0029	5.45E-06	6.24E-05
YK	0.0029	0.0016	5.46E-06	6.24E-05
DD	0.0030	0.0050	5.97E-06	6.64E-05
KW	0.0016	0.0006	7.32E-06	7.91E-05
WN	0.0012	0.0004	9.14E-06	9.62E-05
FD	0.0034	0.0020	1.17E-05	0.0001
PY	0.0026	0.0013	1.48E-05	0.0001
KY	0.0033	0.0019	2.30E-05	0.0002
CK	0.0025	0.0012	2.46E-05	0.0002
ND	0.0024	0.0037	2.67E-05	0.0002
NF	0.0028	0.0016	2.91E-05	0.0003
GC	0.0022	0.0012	3.45E-05	0.0003
FI	0.0034	0.0020	3.73E-05	0.0003
VG	0.0048	0.0031	4.20E-05	0.0004
GK	0.0048	0.0029	4.74E-05	0.0004
SV	0.0042	0.0059	6.71E-05	0.0005
SG	0.0042	0.0060	6.75E-05	0.0005
KF	0.0031	0.0018	7.10E-05	0.0006
MK	0.0024	0.0013	8.06E-05	0.0006
GS	0.0036	0.0052	8.42E-05	0.0006
HH	0.0005	0.0012	9.42E-05	0.0007
VS	0.0040	0.0056	0.0001	0.0008
AY	0.0025	0.0014	0.0001	0.0009
AR	0.0023	0.0036	0.0002	0.0011
WA	0.0011	0.0004	0.0002	0.0012
PP	0.0021	0.0037	0.0002	0.0012
YW	0.0006	0.0002	0.0002	0.0013
NY	0.0034	0.0018	0.0002	0.0014
MP	0.0007	0.0014	0.0002	0.0014
KC	0.0027	0.0013	0.0002	0.0014
TC	0.0022	0.0013	0.0002	0.0014
KV	0.0045	0.0029	0.0002	0.0015
FN	0.0028	0.0018	0.0002	0.0015
LK	0.0066	0.0049	0.0003	0.0016
QS	0.0028	0.0041	0.0003	0.0017
FG	0.0027	0.0016	0.0003	0.0018
CC	0.0019	0.0006	0.0004	0.0020

EP	0.0017	0.0027	0.0004	0.0023
TR	0.0019	0.0029	0.0004	0.0024
EW	0.0011	0.0004	0.0005	0.0025
RP	0.0017	0.0026	0.0006	0.0031
LP	0.0037	0.0050	0.0006	0.0031
PD	0.0020	0.0030	0.0006	0.0032
AW	0.0009	0.0004	0.0009	0.0046
KL	0.0072	0.0057	0.0010	0.0050
AS	0.0037	0.0050	0.0010	0.0050
IV	0.0038	0.0026	0.0010	0.0051
HE	0.0019	0.0011	0.0011	0.0054
ML	0.0025	0.0014	0.0013	0.0065
VV	0.0047	0.0033	0.0015	0.0073
HG	0.0024	0.0015	0.0016	0.0074
NW	0.0010	0.0004	0.0016	0.0074
MV	0.0015	0.0009	0.0016	0.0076
PN	0.0022	0.0031	0.0017	0.0077
KR	0.0035	0.0047	0.0017	0.0077
VC	0.0025	0.0015	0.0018	0.0082
GY	0.0026	0.0016	0.0019	0.0086
KD	0.0038	0.0026	0.0020	0.0087
VW	0.0011	0.0007	0.0020	0.0089
CN	0.0023	0.0013	0.0021	0.0089
NP	0.0021	0.0030	0.0021	0.0089
FK	0.0030	0.0019	0.0023	0.0098
HL	0.0019	0.0028	0.0026	0.0107
IF	0.0026	0.0017	0.0028	0.0117
FT	0.0028	0.0020	0.0032	0.0130
VP	0.0024	0.0034	0.0035	0.0143
DK	0.0043	0.0030	0.0036	0.0143
ET	0.0042	0.0027	0.0039	0.0156
VD	0.0041	0.0030	0.0046	0.0179
IC	0.0017	0.0010	0.0047	0.0183
MY	0.0007	0.0004	0.0049	0.0189
NR	0.0022	0.0029	0.0051	0.0194
FW	0.0005	0.0002	0.0054	0.0205
SH	0.0019	0.0026	0.0055	0.0206
HP	0.0011	0.0016	0.0057	0.0209
WV	0.0007	0.0003	0.0059	0.0215

LR	0.0039	0.0048	0.0060	0.0215
II	0.0045	0.0032	0.0060	0.0215
YI	0.0019	0.0012	0.0063	0.0227
IW	0.0010	0.0005	0.0065	0.0229
AT	0.0027	0.0035	0.0068	0.0237
DH	0.0010	0.0015	0.0068	0.0238
WM	0.0004	0.0001	0.0069	0.0238
YV	0.0024	0.0017	0.0076	0.0259
WP	0.0009	0.0004	0.0076	0.0259
LF	0.0036	0.0026	0.0079	0.0264
PI	0.0020	0.0026	0.0084	0.0279
LI	0.0054	0.0043	0.0098	0.0323
LD	0.0043	0.0053	0.0099	0.0323

Supplementary Table 2.6 Evaluation of 16 classifiers by stratified 10-fold cross-validation on positive training set of *Schistosoma* surface proteins.

Red colour represents higher accuracy and green colour represents lower accuracy.

Machine Learning Technique	Accuracy rounds for 10-fold cross-validation										Overall Accuracy
	1	2	3	4	5	6	7	8	9	10	
Gradient Boosting Machine (GBM)	0.6604	0.7547	0.6981	0.5849	0.6415	0.7547	0.6038	0.6315	0.7115	0.5394	0.6581
RBF SVM	0.5283	0.5283	0.5283	0.5283	0.5283	0.5283	0.5283	0.5192	0.5192	0.5294	0.5266
k-Nearest Neighbors	0.4717	0.4717	0.4717	0.4717	0.4717	0.4717	0.4717	0.4808	0.4808	0.4706	0.4734
Decision Tree	0.6226	0.6981	0.5472	0.6038	0.6038	0.8113	0.5849	0.6923	0.6923	0.5098	0.6366
Random Forest	0.6604	0.7547	0.6981	0.7547	0.6226	0.7170	0.6038	0.5192	0.6731	0.6667	0.6670
Ada Boost	0.5660	0.5094	0.4528	0.7170	0.8113	0.7170	0.5094	0.5385	0.5577	0.5098	0.5889
Gaussian Naive Bayes (GNB)	0.4717	0.4717	0.4717	0.4717	0.4717	0.4717	0.4717	0.4808	0.4808	0.4706	0.4734
Linear Discriminant Analysis (LDA)	0.5283	0.4717	0.4717	0.4717	0.4717	0.4717	0.4717	0.4808	0.4808	0.4706	0.4791
Quadratic Discriminant Analysis (QDA)	0.4717	0.4717	0.4717	0.5283	0.4717	0.4717	0.4717	0.4808	0.4808	0.4706	0.4791
Ridge Regression	0.5283	0.5283	0.5283	0.5283	0.5283	0.5283	0.5283	0.5192	0.5192	0.5294	0.5266
Stochastic Gradient Descent (SGD)	0.4717	0.5094	0.4717	0.4717	0.4717	0.5094	0.4717	0.5192	0.4808	0.4706	0.4848
Perceptron	0.4717	0.4717	0.4717	0.4717	0.4717	0.4906	0.4717	0.4808	0.4808	0.4706	0.4753
Passive Aggressive	0.4717	0.4717	0.4717	0.4717	0.4717	0.4906	0.4717	0.4808	0.4808	0.4706	0.4753
Bernoulli Naive Bayes (BNB)	0.7547	0.8113	0.8302	0.8302	0.7736	0.8679	0.7170	0.8654	0.7308	0.6275	0.7809
Nearest Centroid	0.4717	0.4717	0.4717	0.4717	0.4717	0.4717	0.4717	0.4808	0.4808	0.4706	0.4734
Multi-layer Perceptron (MLP)	0.4717	0.4717	0.4906	0.4717	0.4717	0.4717	0.5283	0.4808	0.4808	0.4706	0.4809

Supplementary Table 2.7 Evaluation of 16 classifiers by stratified 10-fold cross-validation on negative training set of Schistosoma non-surface proteins.

Red colour represents higher accuracy and green colour represents lower accuracy.

Machine Learning Technique	Accuracy rounds for 10-fold cross-validation										Overall Accuracy
	1	2	3	4	5	6	7	8	9	10	
Gradient Boosting Machine (GBM)	0.8936	0.7447	0.7447	0.8085	0.7872	0.8043	0.8043	0.7391	0.8000	0.8889	0.8015
RBF SVM	0.5532	0.5532	0.5532	0.5532	0.5532	0.5652	0.5652	0.5652	0.5556	0.5556	0.5573
k-Nearest Neighbors	0.5532	0.5532	0.4468	0.4468	0.5532	0.5652	0.4348	0.5652	0.5556	0.4444	0.5118
Decision Tree	0.5106	0.5532	0.5319	0.4894	0.4894	0.3913	0.4565	0.5435	0.5333	0.5778	0.5077
Random Forest	0.6383	0.5106	0.5106	0.5745	0.5957	0.6522	0.5217	0.4783	0.5111	0.4667	0.5460
Ada Boost	0.7660	0.7872	0.6809	0.7234	0.5532	0.6739	0.8261	0.6739	0.6000	0.8222	0.7107
Gaussian Naive Bayes (GNB)	0.4468	0.4468	0.4468	0.5532	0.4468	0.4348	0.4348	0.4348	0.4444	0.4444	0.4534
Linear Discriminant Analysis (LDA)	0.4468	0.4468	0.4468	0.4468	0.4681	0.4348	0.4348	0.4348	0.4444	0.4444	0.4449
Quadratic Discriminant Analysis (QDA)	0.5532	0.5532	0.5532	0.5532	0.5532	0.5652	0.5652	0.5652	0.5556	0.5556	0.5573
Ridge Regression	0.4468	0.4468	0.4468	0.4468	0.4468	0.4348	0.4348	0.4348	0.4444	0.4444	0.4427
Stochastic Gradient Descent (SGD)	0.4468	0.4468	0.4468	0.4468	0.4468	0.4348	0.4348	0.4348	0.4444	0.4444	0.4427
Perceptron	0.4468	0.4468	0.4468	0.4468	0.4681	0.4348	0.4348	0.4348	0.6000	0.5111	0.4671
Passive Aggressive	0.4468	0.4468	0.5319	0.5745	0.8298	0.7174	0.6087	0.4783	0.6889	0.6444	0.5967
Bernoulli Naive Bayes (BNB)	0.6809	0.6383	0.7021	0.7872	0.7872	0.6957	0.6739	0.6739	0.7556	0.7333	0.7128
Nearest Centroid	0.4468	0.4468	0.4468	0.4468	0.4468	0.4348	0.4348	0.4348	0.4444	0.4444	0.4427
Multi-layer Perceptron (MLP)	0.4468	0.4468	0.4468	0.4468	0.4468	0.5652	0.5652	0.4348	0.5778	0.4667	0.4844

Chapter 3 Identifying Schistosome-Specific Proteins Immunoreactivity

3.1 Foreword

The following chapter explores a machine learning approach to classify Schistosoma protein immunoreactivity. A modified approach of the method described in chapter 2 used on *Schistosoma* protein microarray data to classify proteins between immunoreactive and non-immunoreactive. SchistoTarget, a machine learning based classifier, have been developed to identify *Schistosoma* proteins immunoreactivity. This chapter describes the method, usage and prediction accuracy of SchistoTarget.

3.2 Abstract

Schistosomiasis is considered by the World Health Organization as the second most socioeconomically devastating and second most common parasitic disease, affecting 200 million people worldwide and causing at least 300,000 deaths annually. No vaccines are available and novel vaccine candidates against schistosomiasis are required. Recently several *Schistosoma* immunomics studies employing protein microarrays have provided essential information for vaccine target identification.

In this project, it is showed that *Schistosoma* proteins recognised by the host immune system have specific sequence properties that can be used to discriminate between immunoreactive and non-reactive proteins. This project results demonstrate that computational predictive methods can likely provide valuable information for the discovery of effective novel vaccine targets. To help prioritizing candidate vaccine targets, the SchistoTarget webserver have been developed, which uses machine learning methods for the identification of *Schistosoma* proteins recognised by the host immune system. The server achieves a sensitivity of 65% and specificity of 72% and provides a user-friendly web-interface. Results are presented in interactive tables and figures. SchistoTarget is publicly available at http://schistotarget.bioapps.org. Source code and documentation are available from https://github.com/shihabhasan/schistotarget.

3.3 Introduction

Schistosomiasis is one of the major neglected tropical diseases (NTDs) causing significant morbidity and mortality of humans residing in tropical countries[58]. Human treatment with praziquantel (PZQ) is used to control schistosomiasis[59] but mass treatment does not prevent reinfection[60]. For a long term disease control there is an urgent need for vaccines, which are not yet available[2]. Recent advances have utilized immunomics approaches in efforts to discover novel vaccine antigens[15,36]. Immunomics provides an invaluable resource for obtaining antibody signatures[15]. Antibody signatures reflect different disease pathologies[61]. The IgE response has been shown to correlate with both allergic reactions and immunity to *Schistosoma*. IgG4 responses are prevalent against allergen-like IgE-binding antigens and IgG1 responses are prevalent against recombinant *S. mansoni* proteins[62].

Schistosoma protein sequence features enable the *in silico* identification of protein class efficiently by using machine learning techniques[27]. Machine learning based *in silico* antigen discovery strategy can lead to effective identification of potentially novel schistosomiasis vaccine antigens.

3.4 Methods

3.4.1 Data set

As training set, host immune response to *Schistosoma* proteins was obtained from a recently published immunonomics study using a protein microarray. A total of 217 protein sequences have been collected of which 215 were RTS (rapid translation system) proteins and 2 are purified recombinant proteins[15]. After removing the isoform proteins 214 sequences remained, where 78 proteins were IgE reactive, 43 were IgG1 reactive, 96 proteins were IgG3 reactive and 21 proteins were IgG4 reactive. Some proteins overlapped between different antibody signatures. After merging these sequences to immonoreactive (recognized by at least one antibody response) and non-immunoreactive (no antibody response recognized) classes, 110 sequences remained as immonoreactive class and 90 sequences remained as non-immunoreactive class. This pilot immunonomics study[15] leads to a relatively small size of the training set.

3.4.2 Features selection

482 features from each protein were extracted. Of these, 82 features represented sequence characteristics and structural and biochemical attributes (Supplementary Table 3.1). The remaining 400 features were 2-mers of amino acids (Supplementary Table 3.2). Features were extracted from each protein sequence using different available bioinformatics tools and in-house Python scripts (Table 3.1). The data followed approximately a normal distribution which was assessed by comparing mean and median values and the shape of the data (

Supplementary Table 3.3). The distributions of features across different antibody types were compared by t-test with a significance level of 0.05.

Table 3.1 Tools used to extract protein features.

Tool	Purpose	URL
Pepstats	Calculation of statistics for proteins such as molecular weight, isoelectric point etc.	http://www.ebi.ac.uk/Tools/seqstats/emboss_pepstats/
Protparam	Computation of various physical and chemical parameters	http://web.expasy.org/protparam/
Garnier	Prediction of protein secondary structure	http://emboss.sourceforge.net/apps/release/6.2/emboss/apps/garnier.html
NetCGlyc	C-mannosylation sites	http://www.cbs.dtu.dk/services/NetCGlyc/
NetChop	Proteasomal cleavages (MHC ligands)	http://www.cbs.dtu.dk/services/NetChop/
NetNGlyc	N-linked glycosylation sites	http://www.cbs.dtu.dk/services/NetNGlyc/
ANCHOR	Prediction of Protein Binding Regions in Disordered Proteins	http://anchor.enzim.hu/
ProP	Arginine and lysine propeptide cleavage sites	http://www.cbs.dtu.dk/services/ProP/
TargetP	Prediction of the subcellular location of eukaryotic proteins	http://www.cbs.dtu.dk/services/TargetP/
BepiPred	Prediction of the location of linear B-cell epitopes	http://www.cbs.dtu.dk/services/BepiPred/
Class I Immunogenecity	Prediction of MHC Class I immunogenicity	http://tools.iedb.org/immunogenicity/
TMHMM	Prediction of transmembrane helices in proteins	http://www.cbs.dtu.dk/services/TMHMM/

3.4.3 Features scaling

The range of values of the different features included in SchistoTarget varies widely. To ensure that each feature contributes approximately proportionately and, therefore, to avoid biases introduced by features with greater numeric ranges[54], all features are scaled into the range of 0 to 1.

3.4.4 Selection of best performing machine-learning technique

The size of the training set is relatively which may be very challenging for machine learning problem. I have applied more machine learning techniques compared to the

previous approach described in Chapter 2 and 21 different machine learning techniques were evaluated on the training set. Classifiers were run using the Scikit-learn (Version 0.18.2) library in Python[55]. Theses 21 classifiers were: (i) Gradient Boosting Machine (GBM), (ii) Ada Boost, (iii) Support Vector Machine with Radial Bias Function kernel (RBF SVM), (iv) Support Vector Machine with Linear kernel (Linear SVM), (v) k-Nearest Neighbors, (vi) Decision Tree, (vii) Random Forest, (viii) Extra Trees Classifier, (ix) Gaussian Naive Bayes (GNB), (x) Multinomial Naive Bayes (MNB), (xi) Bernoulli Naive Bayes (BNB), (xii) Linear Discriminant Analysis (LDA), (xiii) Quadratic Discriminant Analysis (QDA), (xiv) Ridge Regression, (xv) Stochastic Gradient Descent (SGD), (xvi) Perceptron, (xvii) Passive Aggressive, (xviii) Nearest Centroid, (xix) Multi-layer Perceptron (MLP), (xx) Bagging Classifier, (xxi) Gaussian Process Classifier.

Each training sequence was represented by a 67-dimensional feature vector (22 of biochemical and structural features and 45 of 2-mers). The 21 classifiers were evaluated by stratified k-fold (10-fold) cross-validation. In stratified k-fold cross-validation, folds were selected such that the mean response values were approximately equal in all folds[56]. Finally, the mean accuracy was computed for all 10 iterations.

Due to small training data set, a single classifier can provide high false positive prediction rate, it is not feasible to select a single classifier. To reduce the false positive prediction rate, SchistoTarget combined the 2 supervised machine learning techniques which achieved the highest prediction accuracies during the 10-fold cross-validation. The classifiers were combined classifier using a majority-voting rule. A protein is assigned to positive class only if it is predicted by the 2 classifiers as positive; otherwise, SchistoTarget assigns the protein as negative class i.e., only one or no classifiers predict the protein as positive.

3.4.5 Performance evaluation

The classification accuracy of SchistoTarget was evaluated by sensitivity, specificity and overall accuracy[57]. These measures are defined as: sensitivity = TP/(TP+FN), specificity = TN/(TN+FP), overall accuracy = (TP + TN) / (TP + TN + FP + FN), where TP (True Positive) and TN (True Negative) are the number of correctly predicted positive and negative proteins, respectively, and FP (False Positive) and FN (False Negative) are the number of incorrectly predicted positive and negative proteins, respectively. Additionally, the discriminatory power of classifiers was evaluated by the Area Under the Receiver Operating Characteristic (ROC) curve (AUC). Classification performance of SchistoTarget

was assessed on the same training set by leave-one-out cross-validation method (Tables S6 and S7) as independent test data set is not available. Leave-one-out cross-validation is a simple cross-validation method. Each learning set is created by taking all the samples except one, the test set being the sample left out. Thus, for n samples, we have n different training sets and n different tests set. This cross-validation procedure does not waste much data as only one sample is removed from the training set. So, leave-one-out cross-validation is also a k-fold cross-validation where k is equal to the number of samples in the data set[55,63].

3.5 Results

3.5.1 SchistoTarget overview

SchistoTarget incorporates two machine learning techniques (Gaussian Naive Bayes and Bernoulli Naive Bayes) to discriminate between immunoreactive and non-reactive proteins. Generated predictions are stored in a database which facilitates rapid reuse of results without re-running the time-consuming classifiers. This saves considerable runtime if the same sequences are uploaded multiple times, e.g. by different users.

SchistoTarget takes FASTA formatted sequence files or pasted protein sequences as input. If the proteins are already present in the database, the pre-computed results are returned. Otherwise SchistoTarget extracts features from each query sequence using several available bioinformatics tools and newly developed Python scripts. Features include sequence characteristics, biochemical attributes, structural properties and 2-mers. These features are then scaled and used to discriminate between immunoreactive and non-reactive proteins using a majority-voting rule of 2 supervised machine learning techniques (Gaussian Naive Bayes and Bernoulli Naive Bayes). The results of the classification are returned to the user and stored in the database for future reuse. Results are presented as interactive tables, charts and figures. No installation, configuration, registration or login is required. Data are kept privately and automatically deleted and processing has completed.

3.5.2 Discriminating features of *Schistosoma* proteins recognized by different host antibody types system

Associations between 82 biochemical and structural protein features and antibody responses were examined using a t-test. Only a small number of features were significantly differentially distributed ($p<0.05$) (Figure 3.1; Table 3.2) among 4 antibody responses (IgE, IgG1, IgG3 and IgG4). IgE reactive antigens showed a higher frequency of basic mole, and nitrogen sparing than IgG1 reactive antigens. Glutamine frequency was higher in IgG4 than IgE reactive antigens. The percentage of basic mole and arginine and lysine propeptide cleavage sites were less in IgG1 antigens than IgG3 antigens, whereas a higher frequency of grand average of hydropathy (GRAVY) was found in IgG1 antigens. IgG4 antigens had a higher isoelectric point than IgG1 antigens. IgG3 antigens showed a higher frequency of acidic moles, glutamic acid, oxygen sparing, grand average of hydropathy (GRAVY) than IgG4 antigens. Isoelectric point, isoleucine and glutamine frequencies were higher in IgG4 antigen than IgG3 antigens.

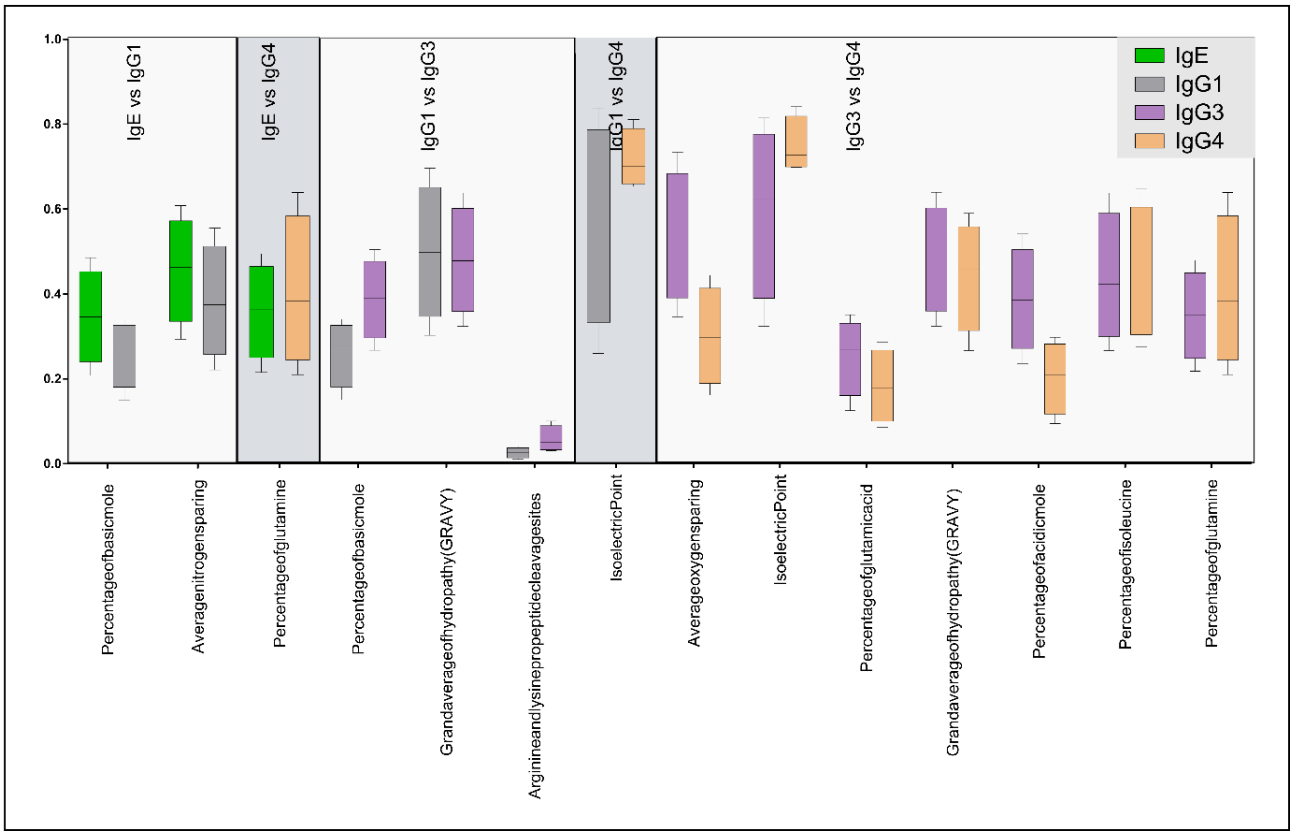


Figure 3.1 The distribution of features among different schistosome antibody signature response proteins.

Means between different antibody signatures were compared by t-test. Only features with $p<0.05$ are shown.

Table 3.2 Features differentially distributed between schistosome antibody signatures.

Means between positive and negative training sets were compared by t-test. All features with p<0.05 are shown.

Antibody Signature Comparison	Features			P-value
IgE vs IgG1		Mean IgE	Mean IgG1	
	Percentage of basic mole	14.3755	12.8767	0.0266
	Average nitrogen sparing	1.3734	1.3428	0.0297
IgE vs IgG4		Mean IgE	Mean IgG4	
	Percentage of glutamine	3.7129	2.7942	0.0438
IgG1 vs IgG3		Mean IgG1	Mean IgG3	
	Percentage of basic mole	12.8767	14.1125	0.0293
	Grand average of hydropathy (GRAVY)	-0.0984	-0.2279	0.0351
	Arginine and lysine propeptide cleavage sites	0.0465	0.1789	0.0479
IgG1 vs IgG4		Mean IgG1	Mean IgG4	
	Isoelectric Point	7.7028	8.6851	0.0302
IgG3 vs IgG4		Mean IgG3	Mean IgG4	
	Average oxygen sparing	2.4759	2.4258	0.0159
	Isoelectric Point	7.8824	8.6851	0.0249
	Percentage of glutamic acid	5.0762	3.7709	0.0308
	Grand average of hydropathy (GRAVY)	-0.2279	-0.0359	0.0314
	Percentage of acidic mole	9.8632	7.9474	0.0316
	Percentage of isoleucine	7.4746	8.7245	0.0317
	Percentage of glutamine	3.7011	2.7942	0.0431

Further, 82 biochemical and structural sequence features for immunoreactive (combined IgE, IgG1, IgG3 and IgG4) and non-immunoreactive schistosome proteins were extracted. Twenty-two features were significantly differentially distributed between immunoreactive and non-immunoreactive proteins (p<0.05) (Figure 3.2; Table 3.3). Immunoreactive proteins showed a higher frequency of cysteine, isoleucine, asparagine, secondary turns, and transmembrane helices. Immunoreactive proteins were also found to be more aromatic and less acidic than non-immunoreactive proteins. Glutamic acid, alanine and secondary helix and were underrepresented in immunoreactive proteins. 45 of 2-mers

were significantly differentially distributed between immunoreactive and non-immunoreactive proteins (p<0.05) (
Supplementary Table 3.4).

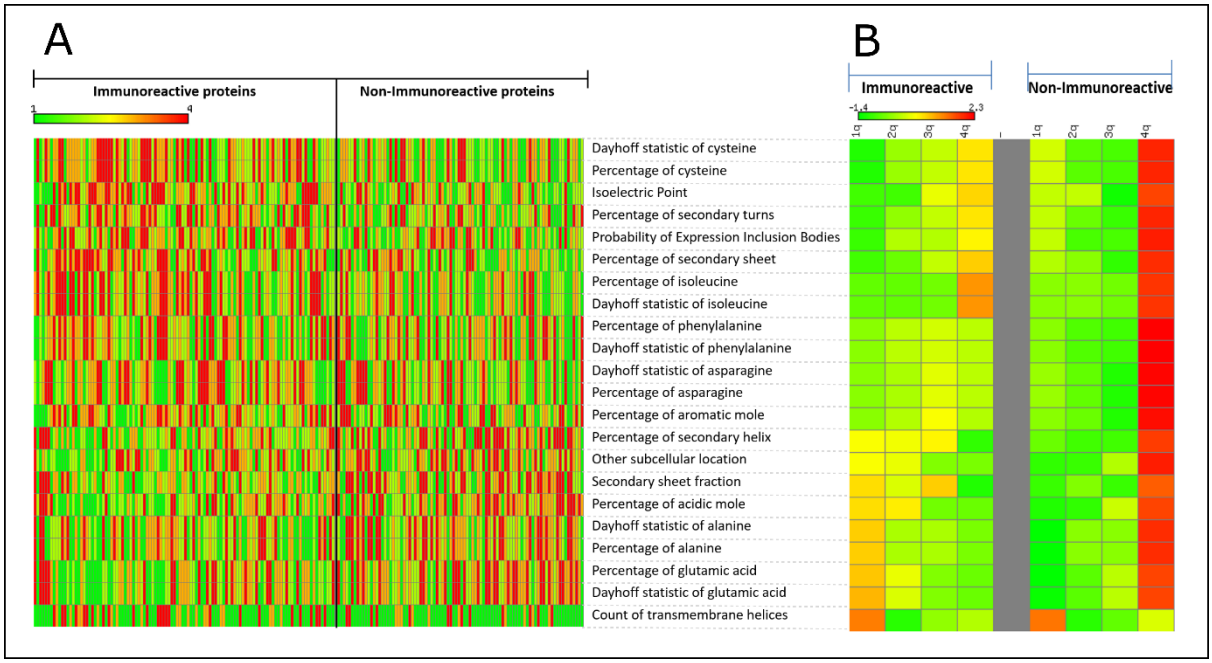


Figure 3.2 Features associated with schistosome immunoreactive proteins.

Means in the immunoreactivity positive and immunoreactivity negative training sets were compared by t-test. All features with p<0.05 are shown. (A) Heatmap of features significantly differentially distributed between immunoreactive and non-immunoreactive proteins. Columns represent each protein of the training set; rows represent features. (B) Quantiles distribution of significantly different features for immunoreactive and non-immunoreactive proteins. Values are depicted in color code, ranging from green (low) to red (high).

Table 3.3 Features differentially distributed between immunoreactive and non-immunoreactive schistosome proteins.

Means between positive and negative training sets were compared by t-test. All features with p<0.05 are shown.

Features	Mean immunoreactive antigens	Mean non-immunoreactive antigens	P-value	False Discovery Rate (FDR)
Dayhoff statistic of glutamic acid	0.8332	1.0177	0.0006	0.0231
Isoelectric Point	7.8531	7.0421	0.0007	0.0231
Percentage of glutamic acid	4.9990	6.1063	0.0008	0.0231
Secondary sheet fraction	0.2156	0.2395	0.0013	0.0262

Percentage of secondary turns	28.1055	23.9200	0.0024	0.0367
Percentage of acidic mole	10.0226	11.4822	0.0027	0.0367
Dayhoff statistic of cysteine	1.0421	0.7919	0.0063	0.0740
Percentage of cysteine	3.0220	2.2966	0.0077	0.0792
Percentage of secondary helix	31.5573	38.3267	0.0089	0.0810
Other subcellular location than mitochondrial or secretory pathway	0.4914	0.6185	0.0145	0.1191
Percentage of secondary sheet	30.5209	26.7700	0.0187	0.1199
Dayhoff statistic of alanine	0.5751	0.6732	0.0215	0.1199
Percentage of alanine	4.9460	5.7901	0.0216	0.1199
Percentage of isoleucine	7.3791	6.4514	0.0219	0.1199
Dayhoff statistic of isoleucine	1.6397	1.4336	0.0219	0.1199
Probability of Expression Inclusion Bodies	0.7549	0.7154	0.0278	0.1427
Count of transmembrane helices	1.1364	0.7556	0.0363	0.1614
Percentage of phenylalanine	4.5930	4.2025	0.0370	0.1614
Dayhoff statistic of phenylalanine	1.2759	1.1673	0.0386	0.1614
Dayhoff statistic of asparagine	1.3330	1.2128	0.0394	0.1614
Percentage of asparagine	5.7317	5.2146	0.0415	0.1621
Percentage of aromatic mole	12.1507	11.3922	0.0464	0.1728

3.5.3 Performance evaluation of 21 machine-learning techniques

The classification accuracy of 21 supervised machine learning techniques was evaluated. Classification performance was assessed on a training set of known *Schistosoma* surface proteins and secreted peptides using stratified k-fold (10-fold) cross-validation (Supplementary Table 3.5). The 21 classifiers achieved classification accuracies in the range of 0.45 - 0.73 (Fig. 3). The combination of the 2 top performing techniques (Gaussian Naive Bayes and Bernoulli Naive Bayes) reduced the false positive prediction rate, achieving a classification accuracy of 0.71 (Figure 3.3;

Table 3.4). The top 2 performing classifiers were therefore incorporated in the SchistoTarget webserver.

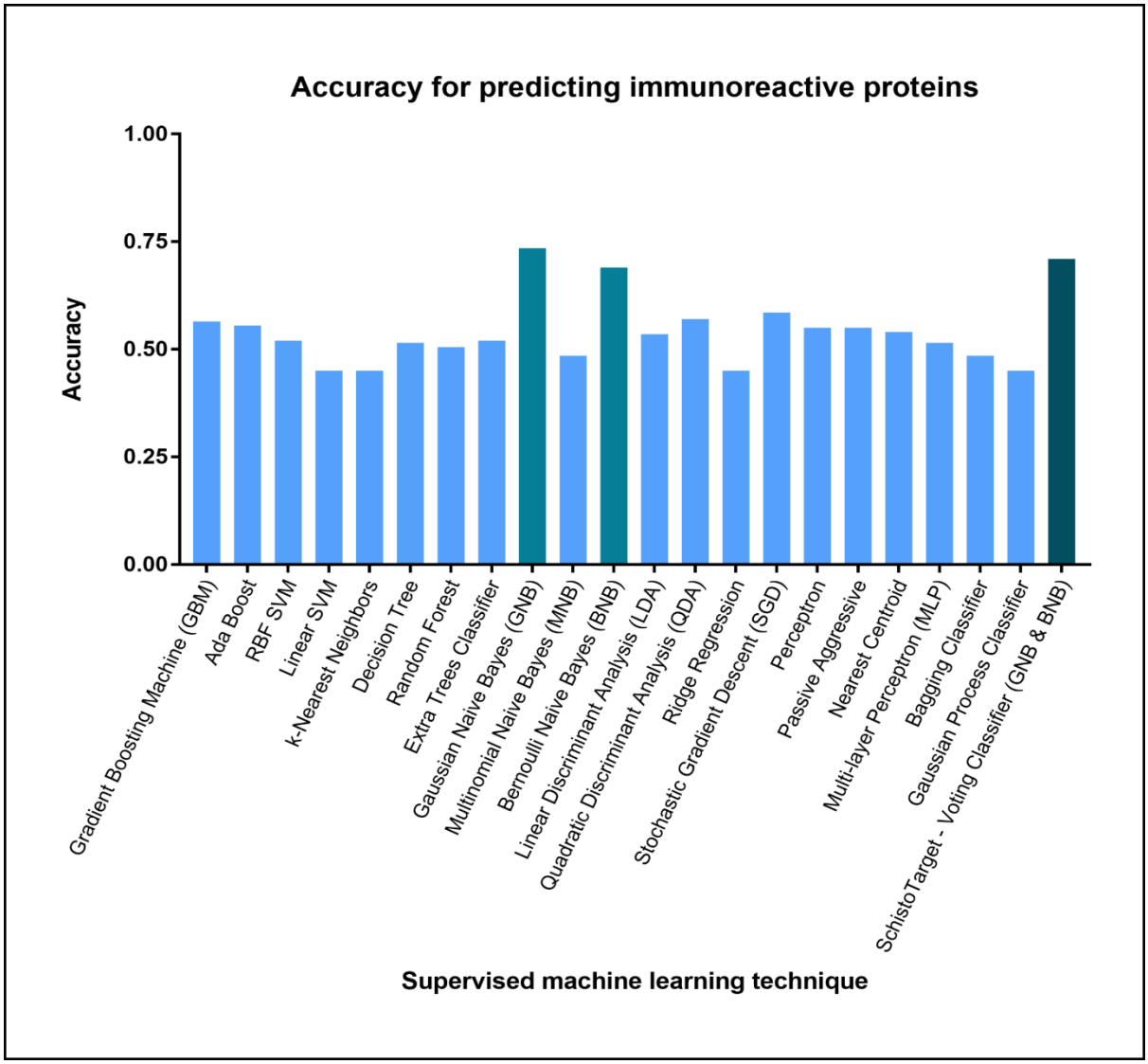


Figure 3.3 Comparison of different supervised machine learning techniques for the identification of Schistosoma immunoreactive and non-immunoreactive proteins.

Classifiers were trained on the training set of known *Schistosoma* immunoreactive and non-immunoreactive proteins and evaluated by stratified k-fold (10-fold) cross-validation.

Table 3.4 Comparison of prediction accuracy of 21 supervised machine learning techniques for Schistosoma immunoreactive proteins.

Classifiers were evaluated on the training set of known immunoreactive (n=110) and non-immunoreactive (n=90) schistosome proteins by stratified k-fold (10-fold) cross-validation. Additionally, the classification accuracy of the SchistoTarget classifier was evaluated, which is based on the combination of the high performing 2 machine learning techniques. AUC: Area Under the Roc Curve.

Machine Learning Technique	Immunoreactivity Classification Overall Accuracy	Surface Classification AUC
Gradient Boosting Machine (GBM)	0.5650	0.5904
Ada Boost	0.5550	0.5510
RBF SVM	0.5200	0.5000
Linear SVM	0.4500	0.5000
k-Nearest Neighbors	0.4500	0.5000
Decision Tree	0.5150	0.5379
Random Forest	0.5050	0.5308
Extra Trees Classifier	0.5200	0.5414
Gaussian Naive Bayes (GNB)	0.7350	0.7187
Multinomial Naive Bayes (MNB)	0.4850	0.5288
Bernoulli Naive Bayes (BNB)	0.6900	0.6859
Linear Discriminant Analysis (LDA)	0.5350	0.5207
Quadratic Discriminant Analysis (QDA)	0.5700	0.5384
Ridge Regression	0.4500	0.5000
Stochastic Gradient Descent (SGD)	0.5850	0.5692
Perceptron	0.5500	0.5556
Passive Aggressive	0.5500	0.5576
Nearest Centroid	0.5400	0.5020
Multi-layer Perceptron (MLP)	0.5150	0.5217
Bagging Classifier	0.4850	0.5126
Gaussian Process Classifier	0.4500	0.5000
SchistoTarget Voting Classifier (GNB & BNB)	0.7100	0.7131

3.5.4 Prediction accuracy of SchistoTarget

SchistoTarget combines 2 supervised machine learning techniques (Gaussian Naive Bayes and Bernoulli Naive Bayes) and classifies proteins based on a majority-voting rule. The performance of SchistoTarget was first evaluated on the training set by stratified 10-fold cross-validation. The final classifier was then evaluated by leave-one-out cross-validation method on the entire data set. SchistoTarget achieved a sensitivity, specificity and overall accuracy of 0.65, 0.72 and 0.69 respectively (Table 3.5;
Supplementary Table 3.6).

Table 3.5 Comparison of prediction accuracy for immunoreactive schistosome proteins.

Prediction accuracy was evaluated on the training set of 110 immunoreactive proteins and 90 non-immunoreactive proteins using the leave-one-out cross-validation method.

Classifier	True Positive	True Negative	False Positive	False Negative	Sensitivity	Specificity	Overall Accuracy
Gaussian Naive Bayes (GNB)	98	51	39	12	0.89	0.57	0.75
Bernoulli Naive Bayes (BNB)	76	56	35	34	0.69	0.62	0.66
SchistoTarget Majority-voting Classifier (GNB & BNB)	72	65	25	38	0.65	0.72	0.69

3.5.5 User-interface and architecture

SchistoTarget provides an easy-to-use graphical user interface (GUI), an extensive help page and user forum. As input, multiple protein sequences can be uploaded or pasted in fasta format. Results are presented in an interactive results page. A table lists the sequence ID of each query sequence, the prediction (immunoreactive/non-immunoreactive) and classification score (number of positive classifiers). A second table lists the individual predictions obtained for each of the 2 classifiers. Additionally, the classification probabilities are shown. The distribution of sequence features in each query protein is presented in a table and in interactive charts and plots (strip chart, heatmap and bar chart).

SchistoTarget is developed in Python using the Django web framework. The server can handle whole-proteome data sets and there is no limit for the number of uploaded query sequences. The server performs background task processing and can process multiple user sessions in parallel. After data submission, a link is provided which gives access to the predictions.

3.6 Conclusion

SchistoTarget is an easy-to-use and fast classifier for the *in silico* identification of *Schistosoma* immunoreactive proteins and their features. However, the size of the training set, which is relatively small with regards to the different antibody responses, could be criticized. This is, however, owed to the currently insufficient data situation. If more data are available in future, it is possible to increase sensitivity, specificity and overall accuracy of SchistoTarget. The software has been optimized for large data sets and allows rapid whole-proteome analysis. SchistoTarget assists researchers in identifying genes important for host-parasite interaction, studying anti-schistosome protective immunity, and identifying candidate vaccine targets. It therefore represents a valuable tool for improving our understanding of *Schistosoma* pathogenicity and host-parasite interaction, and for informing the rational design of much-needed schistosomiasis vaccines.

Supporting information

Supplementary Table 3.1 List of 82 features used in SchistoTarget for protein classification.

SchistoProt uses 482 features for protein classification. Of these 482 features, 82 features represent biochemical and structural properties (shown in this table). The remaining 400 features represent bi-mers of the 20 amino acids (Supplementary Table 3.2).

Percentage of alanine	Secondary sheet fraction	DayhoffStat of threonine
Percentage of cysteine	Average Residue Weight	DayhoffStat of valine
Percentage of aspartic acid	Average carbon sparing	DayhoffStat of tryptophan
Percentage of glutamic acid	Average nitrogen sparing	DayhoffStat of tyrosine
Percentage of phenylalanine	Average sulphur sparing	Percentage of tiny mole
Percentage of glycine	Average oxygen sparing	Percentage of small mole
Percentage of histidine	Average hydrogen sparing	Percentage of aliphatic mole
Percentage of isoleucine	Charge	Percentage of aromatic mole
Percentage of lysine	Molar Extinction Coefficient A280	Percentage of polar mole
Percentage of leucine	Absobance A280	Percentage of non polar mole

Percentage of methionine	Probability of Expression Inclusion Bodies	Percentage of charged mole
Percentage of asparagine	DayhoffStat of alanine	Percentage of acidic mole
Percentage of proline	DayhoffStat of cysteine	Percentage of basic mole
Percentage of glutamine	DayhoffStat of aspartic acid	Percentage of secondary helix
Percentage of arginine	DayhoffStat of glutamic acid	Percentage of secondary sheet
Percentage of serine	DayhoffStat of phenylalanine	Percentage of secondary turns
Percentage of threonine	DayhoffStat of glycine	Percentage of secondary coil
Percentage of valine	DayhoffStat of histidine	C-mannosylation sites
Percentage of tryptophan	DayhoffStat of isoleucine	Proteasomal cleavages (MHC ligands)
Percentage of tyrosine	DayhoffStat of lysine	N-linked glycosylation sites
Molecular Weight	DayhoffStat of leucine	Arginine and lysine propeptide cleavage sites
Aromaticity	DayhoffStat of methionine	Binding Regions in Disordered Proteins
Instability Index	DayhoffStat of asparagine	Mitochondrial targeting peptide (mTP)
Isoelectric Point	DayhoffStat of proline	Secretory pathway signal peptide (SP)
Grand average of hydropathy (GRAVY)	DayhoffStat of glutamine	Other subcellular location
Secondary helix fraction	DayhoffStat of arginine	Linear B-cell epitopes
Secondary turn fraction	DayhoffStat of serine	Class I Immunogenicity Score
Count of transmembrane helices		

Supplementary Table 3.2 List of 400 2-mers used in SchistoTarget for protein classification.

SchistoTarget uses 400 bi-mers of the 20 amino acids for protein classification.

AA	DA	FA	HA	KA	MA	PA	RA	TA	WA
AC	DC	FC	HC	KC	MC	PC	RC	TC	WC
AD	DD	FD	HD	KD	MD	PD	RD	TD	WD
AE	DE	FE	HE	KE	ME	PE	RE	TE	WE
AF	DF	FF	HF	KF	MF	PF	RF	TF	WF
AG	DG	FG	HG	KG	MG	PG	RG	TG	WG
AH	DH	FH	HH	KH	MH	PH	RH	TH	WH
AI	DI	FI	HI	KI	MI	PI	RI	TI	WI
AK	DK	FK	HK	KK	MK	PK	RK	TK	WK
AL	DL	FL	HL	KL	ML	PL	RL	TL	WL

AM	DM	FM	HM	KM	MM	PM	RM	TM	WM
AN	DN	FN	HN	KN	MN	PN	RN	TN	WN
AP	DP	FP	HP	KP	MP	PP	RP	TP	WP
AQ	DQ	FQ	HQ	KQ	MQ	PQ	RQ	TQ	WQ
AR	DR	FR	HR	KR	MR	PR	RR	TR	WR
AS	DS	FS	HS	KS	MS	PS	RS	TS	WS
AT	DT	FT	HT	KT	MT	PT	RT	TT	WT
AV	DV	FV	HV	KV	MV	PV	RV	TV	WV
AW	DW	FW	HW	KW	MW	PW	RW	TW	WW
AY	DY	FY	HY	KY	MY	PY	RY	TY	WY
CA	EA	GA	IA	LA	NA	QA	SA	VA	YA
CC	EC	GC	IC	LC	NC	QC	SC	VC	YC
CD	ED	GD	ID	LD	ND	QD	SD	VD	YD
CE	EE	GE	IE	LE	NE	QE	SE	VE	YE
CF	EF	GF	IF	LF	NF	QF	SF	VF	YF
CG	EG	GG	IG	LG	NG	QG	SG	VG	YG
CH	EH	GH	IH	LH	NH	QH	SH	VH	YH
CI	EI	GI	II	LI	NI	QI	SI	VI	YI
CK	EK	GK	IK	LK	NK	QK	SK	VK	YK
CL	EL	GL	IL	LL	NL	QL	SL	VL	YL
CM	EM	GM	IM	LM	NM	QM	SM	VM	YM
CN	EN	GN	IN	LN	NN	QN	SN	VN	YN
CP	EP	GP	IP	LP	NP	QP	SP	VP	YP
CQ	EQ	GQ	IQ	LQ	NQ	QQ	SQ	VQ	YQ
CR	ER	GR	IR	LR	NR	QR	SR	VR	YR
CS	ES	GS	IS	LS	NS	QS	SS	VS	YS
CT	ET	GT	IT	LT	NT	QT	ST	VT	YT
CV	EV	GV	IV	LV	NV	QV	SV	VV	YV
CW	EW	GW	IW	LW	NW	QW	SW	VW	YW
CY	EY	GY	IY	LY	NY	QY	SY	VY	YY

Supplementary Table 3.3 Normality distribution checking for the extracted data.

Data are almost normally distributed and evaluated by mean, median and shape of the data. Mean and median have almost similar values for a feature with approximately normal shape of the data.

Feature	Immuno Positive					Immuno Negative				
	Mean	Median	Standard Deviation	Kurtosis	Skewness	Mean	Median	Standard Deviation	Kurtosis	Skewness
Percentage of alanine	4.9460	4.4956	2.3568	0.2901	0.8463	5.7901	5.2642	2.5102	0.2602	0.6220
Percentage of cysteine	3.0220	2.4306	2.1133	2.8361	1.6600	2.2966	1.8299	1.6574	2.6131	1.4920
Percentage of aspartic acid	5.0236	4.9948	2.0623	1.8743	0.7225	5.3758	5.4338	1.8780	0.8794	0.1791
Percentage of glutamic acid	4.9990	4.9823	2.7380	2.6950	1.2204	6.1063	5.8917	2.6067	3.9283	1.2635
Percentage of phenylalanine	4.5930	4.1259	2.0381	0.4035	0.8782	4.2025	4.0025	2.0705	0.3701	0.4769
Percentage of glycine	5.4447	5.0862	3.1032	8.0763	2.0363	5.4479	5.0910	2.4012	0.0792	0.3163
Percentage of histidine	2.6242	2.6089	1.4432	0.4540	0.6047	2.6124	2.3285	1.5831	0.9920	0.9871
Percentage of isoleucine	7.3791	6.7308	2.7637	-0.4327	0.5126	6.4514	6.2926	2.4350	2.2771	1.0099
Percentage of lysine	6.5663	6.2163	2.8988	2.0140	1.1418	6.3987	6.3794	2.5442	-0.3862	0.1364
Percentage of leucine	9.4004	9.2548	2.8875	-0.3224	0.3897	9.3486	9.4222	2.7692	0.1017	-0.0001
Percentage of methionine	2.2153	1.9231	1.2054	5.0717	1.4361	2.7059	2.5063	1.4801	2.5599	1.0468
Percentage of asparagine	5.7317	5.4054	2.3095	0.6017	0.8131	5.2146	4.9925	2.2694	-0.5089	0.1316
Percentage of proline	4.2063	3.9270	2.0887	2.0587	0.9692	4.2217	4.1538	2.0758	0.5447	0.5708
Percentage of glutamine	3.6698	3.4583	1.8007	1.4821	0.9847	3.5941	3.3181	1.6891	1.4598	0.9514
Percentage of arginine	4.9328	4.7234	2.2366	0.6487	0.7228	4.8926	5.0618	1.9721	0.4596	0.2966
Percentage of serine	8.4559	8.2968	2.9268	0.3450	0.5354	8.3761	7.8780	3.0811	2.7190	1.1497
Percentage of threonine	5.6593	5.3818	2.3014	5.9142	1.4271	6.0851	5.9140	2.9220	5.4414	1.8769
Percentage of valine	6.1972	6.0142	2.2328	-0.2809	0.4527	6.3023	6.0983	2.0770	0.3521	0.4318
Percentage of tryptophan	1.0581	0.8386	0.8933	0.7123	0.9957	1.0812	0.9653	0.9016	0.4217	0.8970
Percentage of tyrosine	3.8754	3.5099	2.0150	0.3557	0.7145	3.4962	3.3177	1.9057	-0.4719	0.3314
Molecular Weight	30352.3545	24990.2536	24641.2583	28.7604	4.8485	34698.0430	23858.4935	35868.0100	20.7509	4.0724
Aromaticity	0.0953	0.0950	0.0345	0.0331	0.4103	0.0878	0.0926	0.0353	-0.4870	-0.3911
Instability Index	42.0484	41.8757	11.9973	0.3417	0.2284	40.8218	39.3156	12.3232	3.9877	1.2130
Isoelectric Point	7.8531	8.2767	1.6127	-0.9500	-0.5434	7.0421	6.8187	1.5540	-1.1323	0.2454
Grand average of hydropathy (GRAVY)	-0.2133	-0.2887	0.5217	-0.5459	0.4188	-0.2766	-0.3215	0.4265	0.8844	0.7603
Secondary helix fraction	0.3250	0.3134	0.0725	0.7785	0.8200	0.3088	0.3088	0.0603	-0.3885	0.1177
Secondary turn fraction	0.2384	0.2373	0.0456	2.3530	0.7045	0.2326	0.2371	0.0528	1.1719	0.2957
Secondary sheet fraction	0.2156	0.2190	0.0427	0.3213	-0.0785	0.2395	0.2381	0.0532	0.9198	0.3453
Average Residue Weight	113.2841	113.7104	3.3662	1.1479	-0.6265	112.8625	112.7863	3.1426	0.2051	0.3896
Average carbon sparing	5.0553	5.0430	0.2215	0.6251	0.0071	5.0102	5.0192	0.2123	-0.2067	0.1783
Average nitrogen sparing	1.3707	1.3738	0.0866	-0.4734	0.2081	1.3619	1.3678	0.0810	0.7657	0.2418
Average sulphur sparing	0.0524	0.0476	0.0255	2.6940	1.4091	0.0500	0.0447	0.0222	1.2956	0.9765
Average oxygen sparing	2.4744	2.4806	0.0938	-0.5299	0.0163	2.4972	2.4955	0.0745	-0.0476	-0.3603
Average hydrogen sparing	9.9481	9.9653	0.2899	2.0631	-0.3977	9.8919	9.8751	0.2795	1.0549	0.0112
Charge	6.5364	5.7500	10.4574	5.7992	1.2321	1.9111	2.5000	9.5953	10.3497	-1.8698
Molar Extinction Coefficient A280	30297.0000	25440.0000	27181.7115	22.4812	3.8429	33264.2222	20455.0000	34010.2804	6.7368	2.3425
Absobance A280	1.0174	1.0490	0.5002	-0.3825	0.3096	0.9809	0.8675	0.5428	-0.0794	0.5694
Probability of Expression Inclusion	0.7549	0.7690	0.1236	-0.7942	-0.3205	0.7154	0.7055	0.1306	-0.8351	0.2821

Bodies										
DayhoffStat of alanine	0.5751	0.5225	0.2741	0.2908	0.8468	0.6732	0.6120	0.2919	0.2604	0.6226
DayhoffStat of cysteine	1.0421	0.8380	0.7287	2.8371	1.6602	0.7919	0.6310	0.5715	2.6121	1.4920
DayhoffStat of aspartic acid	0.9134	0.9080	0.3750	1.8756	0.7229	0.9774	0.9880	0.3415	0.8802	0.1794
DayhoffStat of glutamic acid	0.8332	0.8300	0.4564	2.6964	1.2207	1.0177	0.9820	0.4345	3.9230	1.2628
DayhoffStat of phenylalanine	1.2759	1.1460	0.5662	0.4041	0.8783	1.1673	1.1115	0.5752	0.3696	0.4770
DayhoffStat of glycine	0.6482	0.6055	0.3694	8.0837	2.0369	0.6485	0.6060	0.2859	0.0787	0.3166
DayhoffStat of histidine	1.3120	1.3045	0.7216	0.4543	0.6048	1.3062	1.1640	0.7916	0.9914	0.9870
DayhoffStat of isoleucine	1.6397	1.4960	0.6142	-0.4327	0.5125	1.4336	1.3985	0.5411	2.2771	1.0102
DayhoffStat of lysine	0.9948	0.9415	0.4392	2.0135	1.1418	0.9695	0.9665	0.3855	-0.3872	0.1355
DayhoffStat of leucine	1.2704	1.2505	0.3902	-0.3218	0.3900	1.2633	1.2735	0.3742	0.1007	-0.0001
DayhoffStat of methionine	1.3031	1.1310	0.7090	5.0732	1.4366	1.5918	1.4745	0.8707	2.5606	1.0470
DayhoffStat of asparagine	1.3330	1.2570	0.5371	0.6024	0.8136	1.2128	1.1610	0.5278	-0.5088	0.1313
DayhoffStat of proline	0.8089	0.7555	0.4017	2.0562	0.9686	0.8119	0.7990	0.3992	0.5449	0.5705
DayhoffStat of glutamine	0.9409	0.8865	0.4617	1.4829	0.9849	0.9216	0.8510	0.4331	1.4590	0.9514
DayhoffStat of arginine	1.0067	0.9640	0.4565	0.6491	0.7230	0.9985	1.0330	0.4025	0.4589	0.2966
DayhoffStat of serine	1.2080	1.1850	0.4181	0.3457	0.5356	1.1966	1.1255	0.4402	2.7188	1.1500
DayhoffStat of threonine	0.9278	0.8820	0.3773	5.9195	1.4278	0.9975	0.9700	0.4790	5.4437	1.8772
DayhoffStat of valine	0.9390	0.9110	0.3383	-0.2808	0.4527	0.9549	0.9240	0.3147	0.3525	0.4319
DayhoffStat of tryptophan	0.8139	0.6450	0.6872	0.7118	0.9956	0.8317	0.7425	0.6935	0.4212	0.8969
DayhoffStat of tyrosine	1.1398	1.0320	0.5927	0.3560	0.7147	1.0283	0.9760	0.5605	-0.4725	0.3312
Percentage of tiny mole	27.5279	27.9580	4.8215	-0.0968	0.2073	27.9957	28.2925	5.0673	3.3005	0.4420
Percentage of small mole	48.6866	48.4270	5.5506	0.7401	0.0657	49.1102	49.3975	5.9865	1.6542	0.0798
Percentage of aliphatic mole	22.9766	21.7800	5.2906	0.5344	0.6711	22.1023	22.0165	4.7267	1.2252	0.3727
Percentage of aromatic mole	12.1507	12.0480	3.5985	-0.2282	0.2444	11.3922	11.0735	4.1381	-0.6118	-0.1578
Percentage of polar mole	47.6625	48.5210	8.9475	-0.8078	-0.2007	48.6557	49.1500	7.4530	0.5382	-0.2236
Percentage of non polar mole	52.3375	51.4790	8.9475	-0.8078	0.2007	51.3443	50.8500	7.4530	0.5382	0.2236
Percentage of charged mole	24.1459	23.8320	6.7123	0.0048	0.3699	25.3858	25.4105	6.1156	-0.0890	-0.0567
Percentage of acidic mole	10.0227	9.6410	3.8062	0.1813	0.5373	11.4822	11.4960	3.5580	0.2016	0.2888
Percentage of basic mole	14.1234	13.8280	4.1306	0.9507	0.7495	13.9037	13.4670	3.5688	-0.1384	0.1823
Percentage of secondary helix	31.5573	31.5500	14.7100	1.1188	0.6554	38.3267	35.9500	18.3979	0.6728	0.7992
Percentage of secondary sheet	30.5209	29.2500	10.7518	-0.4100	0.2397	26.7700	25.5000	10.7425	1.4639	0.8489
Percentage of secondary turns	28.1055	26.2500	10.1855	0.6029	0.7428	23.9200	22.5000	9.0675	0.5007	0.4316
Percentage of secondary coil	18.4218	18.9500	5.9684	-0.5868	-0.0558	20.3978	20.8000	7.8305	4.2234	1.0398
C-mannosylation sites	0.0818	0.0000	0.3354	20.3746	4.4394	0.0222	0.0000	0.1482	42.4083	6.5929
Proteasomal cleavages (MHC ligands)	86.3727	70.5000	74.8790	29.3441	4.9212	100.1778	68.5000	105.0448	20.7688	4.0528
N-linked glycosylation sites	1.6818	1.0000	1.6861	2.4337	1.4043	1.7889	1.0000	2.3726	10.7747	2.7148
Arginine and lysine propeptide cleavage sites	0.1545	0.0000	0.5450	15.4349	3.9018	0.2222	0.0000	0.5359	4.7371	2.3880
Binding Regions in Disordered Proteins	1.5000	0.0000	2.7319	5.4284	2.3414	1.8333	0.0000	3.4649	26.1005	4.2541
Mitochondrial targeting peptide (mTP)	0.1895	0.1300	0.2077	3.6552	1.9481	0.1475	0.0875	0.1587	5.7522	2.2332
Secretory pathway signal peptide (SP)	0.3705	0.1125	0.3949	-1.4161	0.6660	0.2987	0.1075	0.3438	-0.2866	1.1940
Other subcellular location	0.4914	0.5385	0.3361	-1.6682	-0.0716	0.6185	0.7530	0.3110	-0.8473	-0.8272
Linear B-cell epitopes	90.6182	71.5000	99.7642	28.4428	4.5888	114.4556	76.0000	163.3943	39.0369	5.5912
Class I Immunogenicity Score	-0.8027	-0.6406	2.7047	0.7151	-0.3906	-0.4581	-0.7095	3.4638	8.9492	1.4390
Count of transmembrane helices	1.1364	0.0000	1.8939	2.3316	1.7894	0.7556	0.0000	1.8861	16.1960	3.6811

Supplementary Table 3.4 List of 2-mers differentially distributed between immunoreactive and non- immunoreactive schistosome proteins.

Means between positive and negative training sets were compared by t-test. All features with $p<0.05$ are shown.

Features	Mean immunoreactive antigens	Mean non-immunoreactive antigens	P-value
EA	0.0019	0.0045	0.0001
QW	0.0007	0.0001	0.0004
GE	0.0019	0.0036	0.0015
IF	0.0037	0.0018	0.0022
CI	0.0025	0.0012	0.0033
FI	0.0052	0.0030	0.0040
ML	0.0012	0.0025	0.0043
EV	0.0028	0.0044	0.0045
VC	0.0023	0.0011	0.0055
DA	0.0022	0.0037	0.0067
LN	0.0051	0.0035	0.0095
SI	0.0066	0.0049	0.0110
CN	0.0022	0.0010	0.0123
ES	0.0035	0.0053	0.0134
TM	0.0006	0.0018	0.0137
LD	0.0052	0.0033	0.0143
PV	0.0030	0.0019	0.0188
RP	0.0024	0.0014	0.0194
AR	0.0019	0.0031	0.0206
CK	0.0020	0.0011	0.0209
YN	0.0022	0.0014	0.0214
NA	0.0030	0.0021	0.0229
IP	0.0037	0.0025	0.0231
AE	0.0024	0.0039	0.0246
LE	0.0035	0.0048	0.0292
DT	0.0020	0.0031	0.0298
NN	0.0040	0.0027	0.0311
VY	0.0027	0.0017	0.0331
CV	0.0021	0.0012	0.0341

HD	0.0014	0.0008	0.0356
TL	0.0041	0.0058	0.0386
MH	0.0011	0.0005	0.0391
TF	0.0022	0.0034	0.0410
HM	0.0003	0.0008	0.0411
RE	0.0019	0.0030	0.0419
DN	0.0031	0.0022	0.0427
WT	0.0008	0.0003	0.0430
GD	0.0024	0.0037	0.0440
EE	0.0031	0.0045	0.0447
KQ	0.0038	0.0029	0.0450
KH	0.0019	0.0011	0.0465
VE	0.0025	0.0036	0.0473
EG	0.0017	0.0027	0.0478
IY	0.0032	0.0021	0.0481
FD	0.0022	0.0032	0.0484

Supplementary Table 3.5 Comparison of prediction accuracy of 21 supervised machine learning techniques for immunoreactive proteins.

Classifiers were evaluated on the training set of known immunoreactive (n=110) and non-immunoreactive (n=90) proteins by stratified k-fold (10-fold) cross-validation.

Machine Learning Technique	**Accuracy rounds for 10-fold cross-validation**										**Overall Accuracy**
	1	2	3	4	5	6	7	8	9	10	
Gradient Boosting Machine (GBM)	0.5	0.5	0.8	0.65	0.55	0.45	0.65	0.5	0.6	0.45	0.5650
Ada Boost	0.5	0.6	0.6	0.5	0.55	0.45	0.55	0.55	0.7	0.55	0.5550
RBF SVM	0.55	0.55	0.45	0.55	0.55	0.45	0.45	0.55	0.55	0.55	0.5200
Linear SVM	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.4500
k-Nearest Neighbors	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.4500
Decision Tree	0.4	0.5	0.6	0.4	0.55	0.6	0.4	0.45	0.6	0.65	0.5150
Random Forest	0.4	0.5	0.55	0.4	0.45	0.7	0.4	0.5	0.7	0.45	0.5050
Extra Trees Classifier	0.45	0.5	0.55	0.65	0.5	0.55	0.45	0.5	0.6	0.45	0.5200
Gaussian Naive Bayes (GNB)	0.5	0.65	0.7	0.7	0.7	0.9	0.8	0.75	0.95	0.7	0.7350
Multinomial Naive Bayes (MNB)	0.55	0.45	0.55	0.5	0.4	0.5	0.45	0.45	0.5	0.5	0.4850
Bernoulli Naive Bayes (BNB)	0.55	0.7	0.65	0.85	0.65	0.65	0.95	0.55	0.65	0.7	0.6900
Linear Discriminant Analysis (LDA)	0.45	0.8	0.55	0.5	0.5	0.5	0.45	0.5	0.5	0.6	0.5350
Quadratic Discriminant Analysis (QDA)	0.55	0.55	0.6	0.45	0.55	0.7	0.55	0.65	0.55	0.55	0.5700
Ridge Regression	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.4500
Stochastic Gradient Descent (SGD)	0.45	0.6	0.45	0.6	0.65	0.45	0.75	0.5	0.65	0.75	0.5850
Perceptron	0.35	0.6	0.5	0.6	0.55	0.5	0.55	0.45	0.75	0.65	0.5500
Passive Aggressive	0.5	0.55	0.6	0.45	0.5	0.55	0.7	0.4	0.6	0.65	0.5500

Nearest Centroid	0.55	0.45	0.55	0.55	0.55	0.55	0.55	0.55	0.55	0.55	0.5400
Multi-layer Perceptron (MLP)	0.55	0.5	0.5	0.45	0.45	0.75	0.5	0.6	0.35	0.5	0.5150
Bagging Classifier	0.4	0.35	0.55	0.65	0.6	0.6	0.5	0.4	0.4	0.4	0.4850
Gaussian Process Classifier	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.4500
Voting Classifier (GNB & BNB)	0.55	0.65	0.75	0.85	0.65	0.7	0.95	0.6	0.7	0.7	0.7100

Supplementary Table 3.6 Comparison of prediction accuracy for immunoreactive proteins.

Prediction accuracy was evaluated on the training set of 110 immunoreactive proteins and 90 non-immunoreactive proteins using leave-one-out cross-validation method.

Immunoreactive Proteins						Non-immunoreactive Proteins					
Round	Sequence ID	Expected label	Predicted GNB	Predicted BNB	Majority-Voting	Round	Sequence ID	Expected label	Predicted GNB	Predicted BNB	Majority-Voting
1	Sj_AY815690	1	0	0	0	111	Sj_AY808749	0	1	1	2
2	Sj_AY811988	1	0	1	1	112	Sj_AY808751	0	1	1	2
3	Sj_AY813185	1	0	0	0	113	Sj_AY222874	0	1	1	2
4	Sj_AY811797	1	0	0	0	114	Smp_046290	0	0	1	1
5	Sj_AY812161	1	1	1	2	115	Sj_AY813118	0	1	1	2
6	Sj_AY815838	1	1	1	2	116	Sj_AY811628	0	0	0	0
7	Sj_AY809620	1	1	1	2	117	Sj_AY809406	0	1	1	2
8	Sj_AY813602	1	1	1	2	118	Sj_AY915861	0	0	0	0
9	Sj_AY222951	1	1	1	2	119	Sj_AY813275	0	1	0	1
10	Sj_AY810792	1	1	1	2	120	Sj_AY915571	0	1	1	2
11	Sj_EF553319	1	1	1	2	121	Sj_AY809115	0	1	1	2
12	Sj_AY816000	1	1	1	2	122	Sj_AY915907	0	0	0	0
13	Sj_AY810537	1	1	0	1	123	Sj_AY815649	0	0	1	1
14	Sj_AY814261	1	1	1	2	124	Sj_AY813876	0	0	0	0
15	Sj_AY814497	1	1	0	1	125	Smp_012440	0	1	1	2
16	Sj_AY815303	1	1	1	2	126	Sj_AY915878	0	0	0	0
17	Sj_AY222868	1	1	1	2	127	Smp_131910	0	1	1	2
18	Sj_AY809911	1	1	0	1	128	Sj_AY812720	0	0	0	0
19	Sj_AY815056	1	0	1	1	129	Sj_AY915721	0	0	0	0
20	Sj_AY810700	1	1	1	2	130	Sj_AY811479	0	1	1	2
21	Sj_AY812195	1	0	0	0	131	Smp_141680	0	1	0	1
22	Sj_AY815945	1	1	1	2	132	Sj_AY811014	0	0	1	1
23	Sj_AY814817	1	1	1	2	133	Sj_AY809555	0	1	1	2
24	Sj_AY814738	1	1	1	2	134	Sj_AY814007	0	1	1	2
25	Sm29	1	1	1	2	135	Sj_AY815489	0	1	1	2
26	SmTSP2	1	1	1	2	136	Sj_AY815616	0	0	1	1
27	Smp_139970	1	1	1	2	137	Smp_130300	0	0	0	0
28	Sj_AY812458	1	1	1	2	138	Sj_AY915793	0	0	1	1
29	Smp_124240	1	1	1	2	139	Smp_145290	0	1	0	1
30	Smp_056970.1	1	0	0	0	140	Sj_AY223001	0	1	0	1

31	Sj_AY814116	1	1	1	2	141	Smp_077720	0	0	0	0
32	Sj_AY809550	1	1	1	2	142	Sj_AY808756	0	0	0	0
33	Sj_AY813467	1	1	0	1	143	Sj_AY814468	0	1	0	1
34	Sj_AY814537	1	1	1	2	144	Sj_AY816005	0	1	1	2
35	Sj_AY808785	1	1	0	1	145	Smp_045500	0	1	0	1
36	Sj_AY808827	1	1	1	2	146	Sj_AY812565	0	0	0	0
37	Sj_AY812976	1	0	1	1	147	Sj_AY814600	0	0	0	0
38	Sj_AY809019	1	1	1	2	148	Sj_AY809244	0	1	1	2
39	Smp_156590	1	1	0	1	149	Smp_121950	0	1	0	1
40	Smp_050270	1	1	1	2	150	Smp_030920	0	0	0	0
41	Sj_AY812470	1	1	1	2	151	Sj_AY810132	0	1	1	2
42	Sj_AY815442	1	1	1	2	152	Sj_AY813942	0	1	1	2
43	Smp_136640	1	1	1	2	153	Sj_AY809972	0	0	0	0
44	Sj_AY814430	1	1	1	2	154	Sj_AY813221	0	0	1	1
45	Sj_AY808953	1	1	1	2	155	Sj_AY226984	0	1	1	2
46	Smp_008310	1	1	1	2	156	Sj_AY814115	0	0	0	0
47	Sj_AF036955	1	1	1	2	157	Sj_AY809388	0	0	0	0
48	Sj_AY813455	1	1	1	2	158	Sj_AY811902	0	0	0	0
49	Sj_AY815815	1	0	1	1	159	Smp_042020	0	0	1	1
50	Smp_147140	1	1	1	2	160	Smp_000100	0	0	0	0
51	Smp_003990	1	1	0	1	161	Sj_L23322	0	0	0	0
52	Smp_008660.1	1	1	0	1	162	Sj_AY814882	0	0	0	0
53	Sj_AY813641	1	1	1	2	163	Sj_AY815177	0	0	0	0
54	Sj_AY812951	1	1	1	2	164	Smp_151490	0	0	0	0
55	Sj_AY809028	1	1	1	2	165	Sj_AY814107	0	0	0	0
56	Smp_096760	1	1	0	1	166	Smp_101970	0	0	0	0
57	Sj_AY812977	1	1	1	2	167	Sj_AY223437	0	1	0	1
58	Sj_AY812972	1	1	0	1	168	Sj_AY915388	0	1	0	1
59	Smp_002880.1	1	0	0	0	169	Sj_AF048759	0	0	0	0
60	Sj_AY814534	1	1	1	2	170	Sj_AY812897	0	0	0	0
61	Sj_AY815248	1	1	0	1	171	Sj_M63706	0	1	0	1
62	Sj_AY816003	1	1	1	2	172	Smp_037540.2	0	1	0	1
63	Sj_AY811126	1	1	1	2	173	Sj_L08198	0	1	0	1
64	Sj_AY810129	1	1	0	1	174	Sj_AF380366	0	0	0	0
65	Sj_AY815196	1	1	1	2	175	Sj_AY815038	0	1	1	2
66	Sj_AY809768	1	1	0	1	176	Sj_AY808393	0	0	0	0
67	Sj_AY816125	1	1	1	2	177	Sj_AY813732	0	1	1	2
68	Sj_AY808893	1	1	1	2	178	Smp_095360.3	0	0	0	0
69	Sj_AY808459	1	1	1	2	179	Smp_153390.2	0	1	1	2
70	Sj_AY533028	1	1	1	2	180	Sj_AY812989	0	0	0	0
71	Sj_AF072327.1	1	1	1	2	181	Smp_017430	0	1	1	2
72	Sj_AY815419	1	1	1	2	182	Sj_AY810680	0	0	1	1
73	Smp_045200	1	1	0	1	183	Sj_AY808379	0	1	1	2
74	Sj_AY808903	1	1	1	2	184	Sj_AY812658	0	0	0	0
75	Smp_075420	1	1	1	2	185	Sj_AY813104	0	0	0	0
76	Sj_AY810692	1	1	1	2	186	Sj_AY816048	0	0	0	0

77	Sj_AY812591	1	1	1	2
78	Sj_AY808650	1	1	1	2
79	Sj_AY814158	1	1	0	1
80	Smp_124050.4	1	1	1	2
81	Sj_AY815101	1	1	1	2
82	Sj_AY223099	1	1	0	1
83	Sj_AY810722	1	1	1	2
84	Sj_AY813439	1	1	1	2
85	Sj_AY815834	1	1	0	1
86	Sj_AY223465	1	0	0	0
87	Sj_AY816044	1	1	1	2
88	Sj_AY814977	1	1	1	2
89	Smp_176200.2	1	1	1	2
90	Sj_AY808899	1	1	0	1
91	Sj_AY222926	1	1	0	1
92	Sj_AY812444	1	1	1	2
93	Sj_AY808494	1	1	0	1
94	Sj_AY814201	1	1	1	2
95	Sj_AY810705	1	1	0	1
96	Sj_AY809338	1	1	1	2
97	Sj_AY814549	1	1	0	1
98	Smp_140000	1	1	1	2
99	Sj_AY814773	1	1	1	2
100	Sj_AY808797	1	1	0	1
101	Sj_AY811460	1	0	0	0
102	Smp_194970	1	1	0	1
103	Sj_AY809286	1	1	1	2
104	Sj_AY814310	1	1	1	2
105	Sj_AY808337	1	1	1	2
106	Smp_004470.2	1	1	0	1
107	Smp_005740	1	1	1	2
108	Sj_AY809526	1	1	1	2
109	Smp_040680	1	1	1	2
110	Smp_151480	1	1	0	1
187	Smp_137410	0	0	0	0
188	Sj_AY813229	0	0	0	0
189	Smp_137170	0	0	0	0
190	Sj_AY810377	0	0	1	1
191	Sj_AY813612	0	0	0	0
192	Sj_AY813810	0	0	0	0
193	Sj_M14654	0	0	0	0
194	Sj_AY809239	0	0	0	0
195	Sj_AY815164	0	1	0	1
196	Sj_AY808531	0	0	0	0
197	Smp_059480	0	1	1	2
198	Sj_AY814401	0	0	0	0
199	Sj_AY813596	0	1	0	1
200	Sj_AY815791	0	1	1	2

Chapter 4 Identifying Putative Drug and Vaccine Targets Against Schistosomiasis

4.1 Foreword

This chapter describes a comparative analysis of Schistosoma genomes and an integrative bioinformatics pipeline to identify putative vaccine targets against schistosomiasis. A set of genes were selected by comparative analysis of several parasite genomes, then these genes were annotated using the developed tools described in chapters 2 and 3. Potential antigens as drug and vaccine targets were selected using Gene Ontology and Swiss-Prot annotations. Finally, protein-protein and protein-chemical interactions were explored using STRING and STICH.

4.2 Abstract

In addition to providing a unique resource for studying evolutionary processes, *Schistosoma* genomes can be used to identify genes important for host-parasite interactions and to discover novel vaccine and drug targets. Conventional approaches for anti-schistosomiasis vaccine development have focused on a limited number of antigens. Recently whole genome sequence data for the three main schistosome species infecting humans (*S. mansoni, S. haematobium* and *S. japonicum*) became available. These datasets provide a unique foundation for a novel approach to anti-schistosomiasis vaccine development. Here in this project, putatively important protective schistosome antigens have been identified from genomic-based information using novel Bioinformatics methods in comparative analysis of the genomes of the three schistosomes infecting humans*, Schistosoma bovis,* which infects ruminants, and the related, but free-living flatworm, *Schmidtea mediterranea*. 345 core genes were identified which are present in all three human-infecting schistosome genomes but absent in *S. bovis* and *S. mediterranea.* Further, targeting immunogenic surface and secretory proteins 20 proteins as potential vaccine targets have been selected. These potential vaccine targets were then *in silico* characterized using Bioinformatics methods to indicate their biological relevance. These putative vaccine targets can be biologically validated by wet laboratory experiments in animals. The Python scripts, used for the analysis, are available from https://github.com/shihabhasan/schistocomp.

4.3 Introduction

Blood flukes of the genus *Schistosoma* (phylum Platyhelminthes) are the cause of schistosomiasis, a chronic disease and a major health concern in Africa and the Asia Pacific Region and Africa. It is considered by the World Health Organization as the second most socioeconomically devastating and second most common parasitic disease[1,2], causing at least 300,000 deaths annually[3]. Treatment relies mainly on a single drug, praziquantel, which does not prevent re-infection and there is a constant concern that drug resistance might develop[2]. Three main *Schistosoma* species infect humans: *S. mansoni* and *S. japonicum,* cause intestinal/hepatic schistosomiasis whereas *S. haematobium* results in urinogenital disease[59]. Conventional approaches, focusing on a very limited number of antigens for anti-schistosomiasis vaccine development, have thus far failed[59]. Driven by the need to improve treatment and prevent infection, the genomes of these three schistosomes have recently become publicly available[20-22]. These genomic-based datasets provide a unique resource for a novel approach to schistosomiasis vaccine development but it has not been clear how this information can be used to identify key antigens as vaccine targets.

Schistosome tegumental surface proteins are responsible for essential functions crucial for parasite survival[23] and secretory peptides modulate host immune responses[40]. During the past decade, schistosome surface and secretory proteins have been considered sources of putative vaccine antigens. Recently, immunomics approaches have been utilized successfully for vaccine antigen discovery[15,36].

I hypothesise that protein-encoding genes present in *S. mansoni, S. haematobium* and *S. japonicum* but absent in the genomes of *Schistosoma bovis,* which infects ruminants, and *Schmidtea mediterranea,* a free-living flatworm phylogenetically related to schistosomes[64], might provide new insight on suitable candidates as schistosomiasis vaccine targets (Figure 4.1).

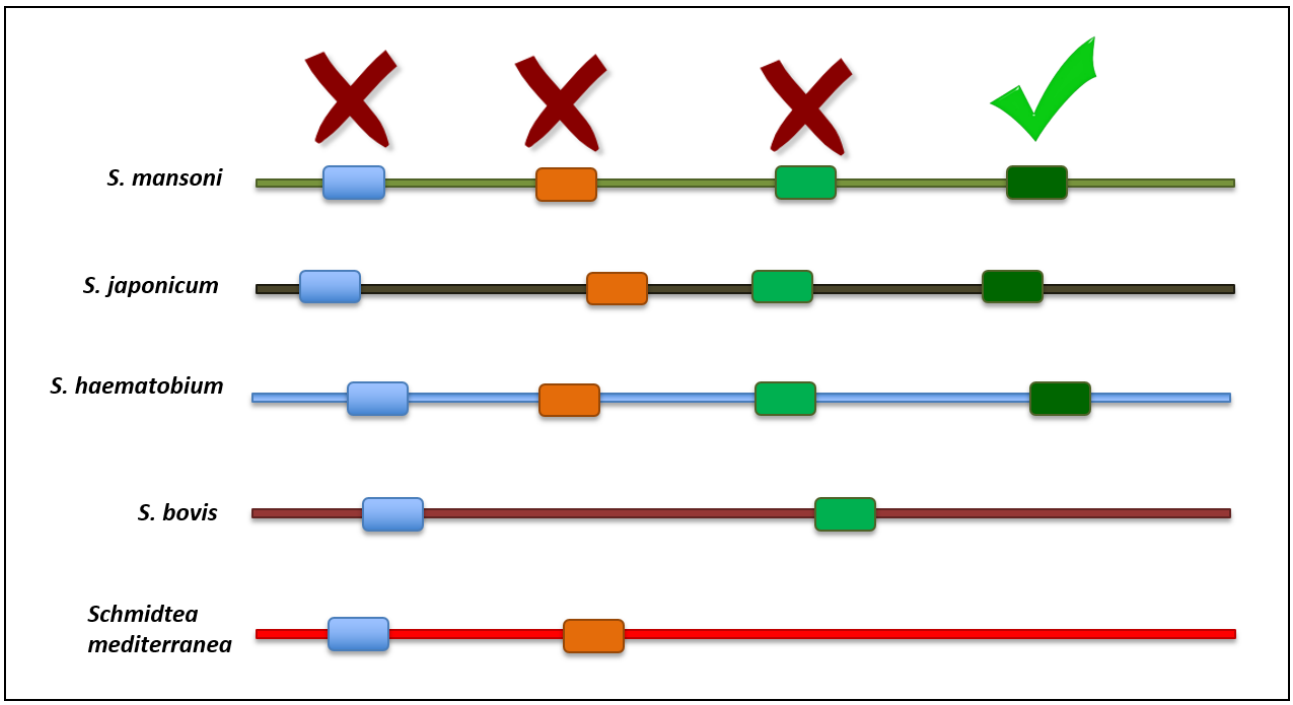


Figure 4.1 Hypothesis to identify putative vaccine targets.

Genes present in human-infecting genomes but absent animal and non-parasitic free-living flatworm might be potential antigens for vaccine targets. *S. mansoni, S. haematobium* and *S. japonicum* infect human. *Schistosoma bovis* infects ruminants and *Schmidtea mediterranea* is a non-parasitic, free-living flatworm.

4.4 Methods

4.4.1 Data

Protein sequences for the three human infecting schistosomes were collected from the SchistoDB[65] database. The *S. mansoni* proteome contains 11,774 proteins, the *S. haematobium* (Egyptian strain) proteome has 11,140 proteins and the *S. japonicum* (Anhui strain) proteome comprises 12,657 proteins. The *S. bovis* genome has been sequenced and predicted 12,924 proteins for a project at QIMR Berghofer Medical Research Institute, Australia. The *Schmidtea mediterranea* proteome sequences were collected from the WormBase ParaSite[66] database which contains 29,850 proteins. The Gene Ontology[67] (GO) annotations for *S. mansoni,* extracted (on July 05, 2017) from the GO Consortium annotation[68] using AmiGO[69] , contains 25,959 annotations. 13,517 *S.mansoni* protein annotations were extracted from Swiss-Prot[70].

4.4.2 Orthologous/core genes prediction

I used the Reciprocal Best Hits (RBH) method incorporating NCBI BLAST[71,72] to identify orthologous proteins among the three human infecting *Schistosoma* species. RBH is found when proteins from different organisms that are each other's top BLAST hit, each in a different genome, when the proteomes from those organisms are compared to each other genomes[73].

The steps for RBH are: i) Take two FASTA files (species A and species B), ii) Build a BLAST database for each, iii) Run reciprocal BLAST searches (A vs B, and B vs A), iv) Filter the High-scoring Segment Pairs (HSPs), and v) Then compile a list of the reciprocal best hits (RBH). The filter E-value of $1e^{-5}$, minimum percentage identity for BLAST matches of 70% and minimum percentage query coverage for BLAST matches of 70% as the best scoring match for BLAST searching were used (Figure 4.2). The RBH was performed for two species at a time among all the *Schistosoma* species and the core genes were identified. The human-infecting *Schistosoma* core genes present in *S. bovis* and *Schmidtea mediterranea* by the RBH method were identified and excluded them as genes of interest.

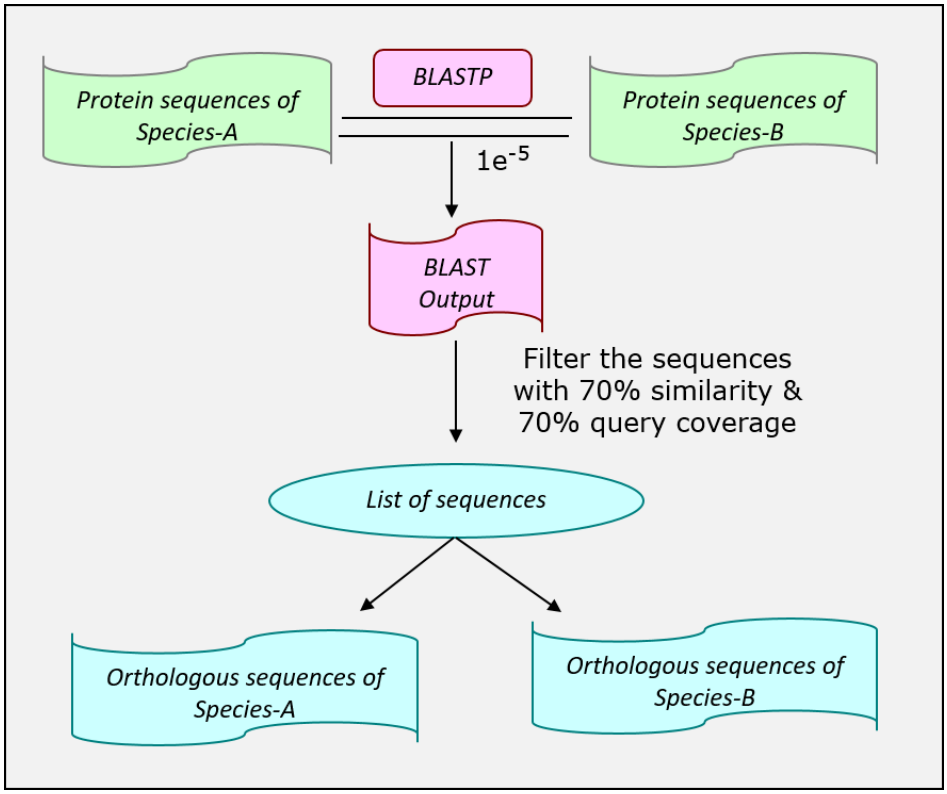


Figure 4.2 RBH method to select orthologous proteins from two different genomes.

RBH runs reciprocal BLAST searches against the proteomes of two species (species A vs species, and species B vs species A). The best scoring match is selected if the High-scoring Segment Pairs (HSPs) have at least 70% identity, 70% alignment length and $1e^{-5}$ E-value.

4.4.3 Protein annotation

Surface and secretory proteins were predicted and selected using SchistoProt[74]. Then, Schistotarget[75] was used to predict which of the surface or secretory molecules were immunoreactive proteins. The immunoreactive proteins were further annotated using the GO[68] and Swiss-Prot[70] data available for *S. mansoni.* GO Enrichment Analysis for the selected proteins was performed using PANTHER (protein annotation through evolutionary relationship) Classification System[76,77]. The protein-protein direct (physical) interactions, as well as indirect (functional) interactions were predicted using STRING[78]. Interactions between proteins and chemicals were predicted using STITCH[79].

4.5 Results

4.5.1 Vaccine Target Identification

An integrative bioinformatics pipeline was employed to identify schistosome vaccine targets (Figure 4.3).

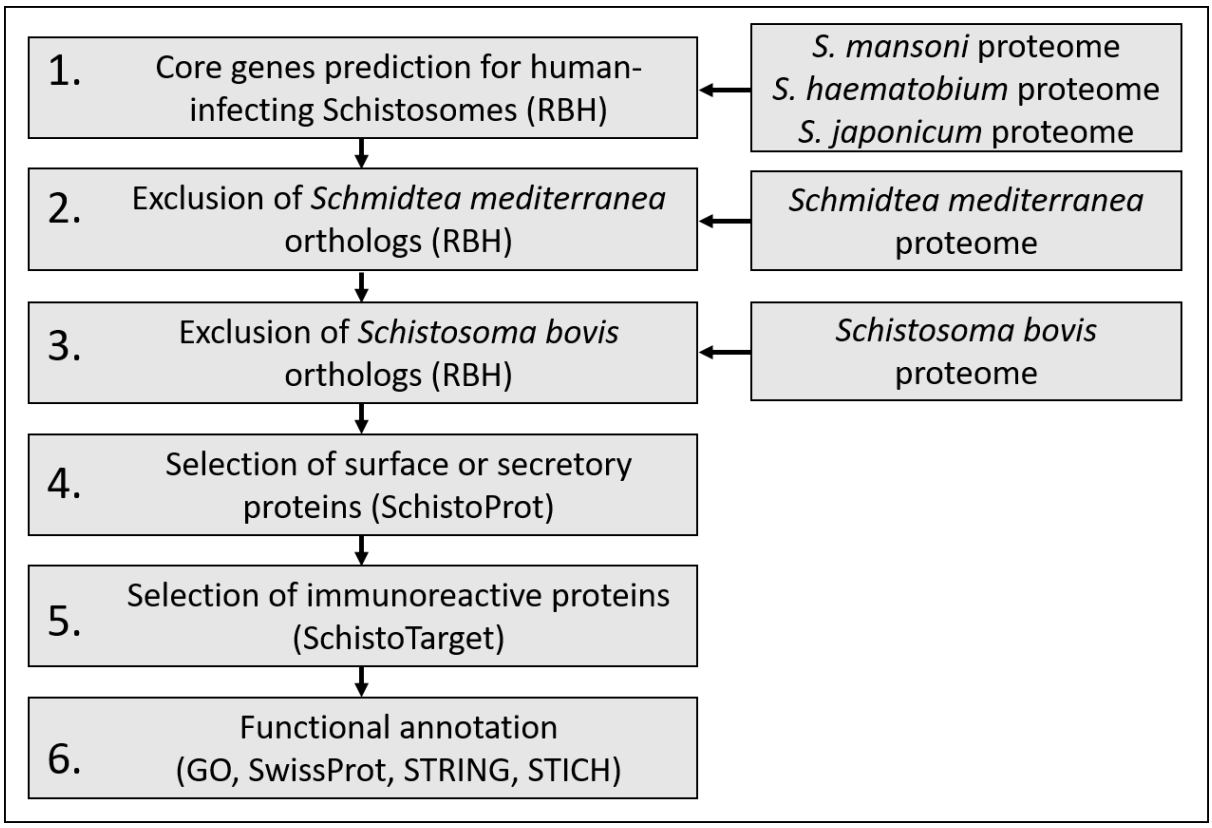


Figure 4.3 Bioinformatics pipeline used to characterize and curate putative schistosome vaccine targets.

In step1, core genes in the human-infecting *S. mansoni*, *S. haematobium* and *S. japonicum* genomes were predicted. In steps 2–3, orthologs from the free-living flatworm *Schmidtea mediterranea* and S. bovis were removed from the core genes from the three human-infecting schistosomes. In steps 4 and 5, surface and secretory proteins were predicted using SchistoProt; protein immunoreactivity was predicted using SchiatoTarget. In step 6, all possible vaccine targets were functionally annotated using GO, SwissProt, STRING and STICH.

First, 6,016 orthologous proteins between *S. mansoni* and S. *haematobium*, 4,209 orthologous proteins between *S. mansoni* and S. *japonicum* and 4,305 orthologous proteins between S. *haematobium* and S. *japonicum* we identified by BLAST RBH. We identified 2,701 core proteins in these three genomes. Then, 177 orthologous proteins were identified between the genomes of these three schistosome species and *Schmidtea mediterranea.* After removing these 177 orthologs from the core proteins, 2,524 proteins remained. 2,179 of the 2,524 proteins were also present in the *S. bovis* proteome. Finally, 345 proteins (Figure 4.4; Supplementary Table 4.1), which remained after removing these orthologs from the 2,524 proteins selected in the previous step, were explored further.

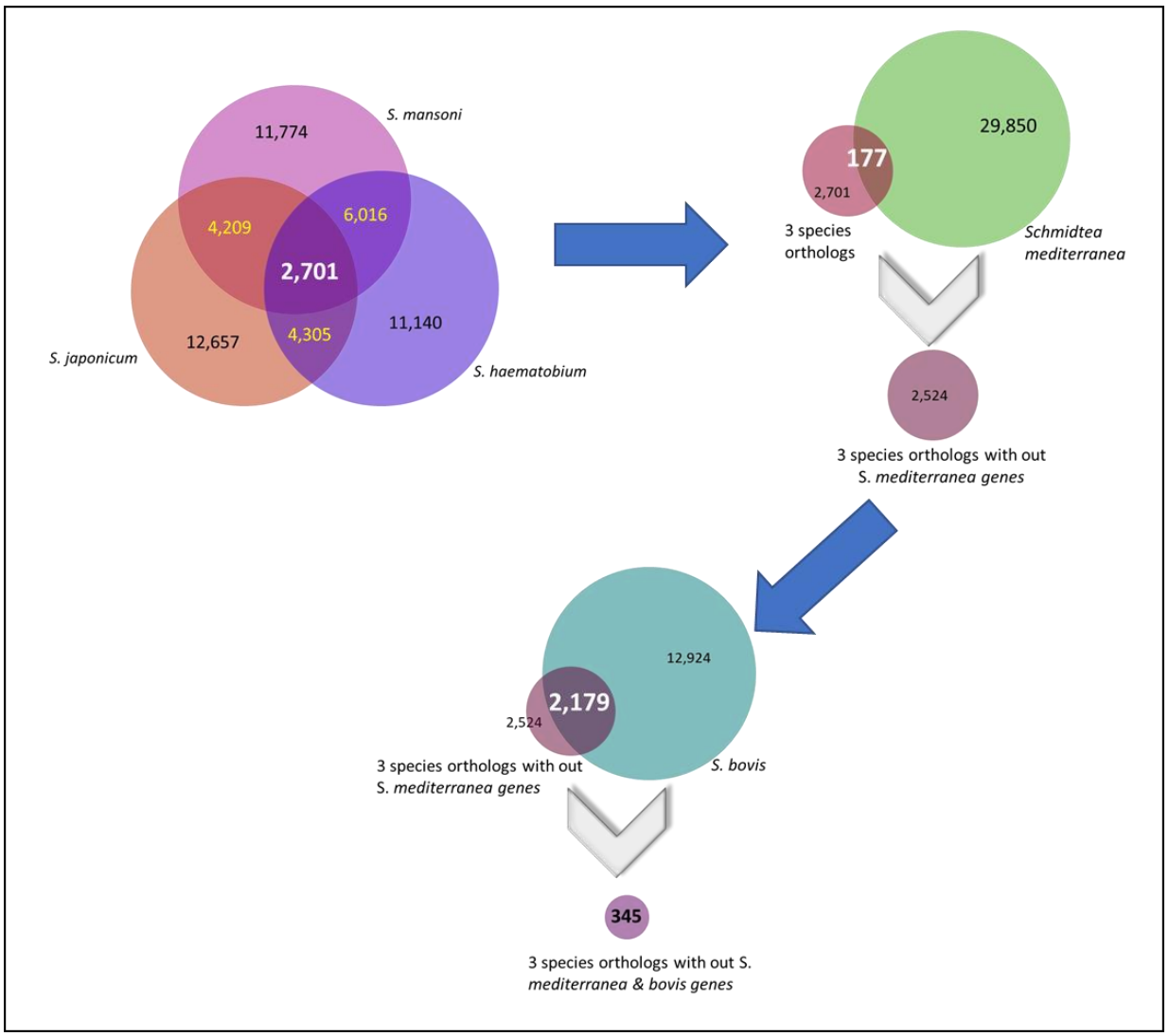


Figure 4.4 Steps involved in the selection of potential vaccine targets using proteomes from different flatworm species.

Step-1 uses RBH against the proteomes from the three human-infecting *Schistosoma* spp.to predict core genes. *S. mediterranea* orthologs were excluded from the core genes in step-2. In step-3, *S. bovis* orthologs were excluded from the genes obtained in step-2.

4.5.2 Prediction of surface, secretory and immunoreactive proteins

83 surface proteins and 106 secretory proteins were predicted from the 345 identified proteins. After merging both surface and secretory proteins, 135 proteins (Supplementary Table 4.2) remained, of which 45 were predicted to be immunreactive (Supplementary Table 4.3).

4.5.3 Protein annotation

44 proteins were mapped to the GO database. 20 proteins (Table 4.1; Supplementary Table 4.4) were selected using GO annotation with biological processes and molecular functions, which are important for host-parasite interactions such as catalytic activity, transmembrane transporter activity, lipid transporter activity, serine-type peptidase activity, serine protease inhibitory activity, G-protein coupled receptor activity and oxidoreductase activity[80-86]. These 20 proteins were further annotated using Swiss-Prot (Table 4.1).

Table 4.1 20 protein antigens, and their annotation, identified as potential schistosomiasis vaccine targets.

Proteins were selected based on comparative genomics and GO annotation.

S. mansoni **gene ID**	***S. mansoni*** **UniProt Gene Symbols**	***S. japonicum*** **ortholog**	***S. haematobium*** **ortholog**	**Annotation**
Smp_002870	G4VLJ1	Sjp_0089930	MS3_07466	G-protein modulator
Smp_017620	G4VGX8	Sjp_0007190	MS3_05025	Amine oxidase
Smp_018990	G4VG38	Sjp_0075950	MS3_04781	60s ribosomal protein L9
Smp_048540	G4VSB9	Sjp_0060110	MS3_05665	Dolichol kinase
Smp_054010	G4VHR0	Sjp_0079140	MS3_03571	Cationic amino acid transporter
Smp_083990	G4LVW7	Sjp_0048940	MS3_02574	Cationic amino acid transporter
Smp_124020	G4V6N7	Sjp_0094660	MS3_09149	Heparan sulfate 6-o-sulfotransferase
Smp_132080	G4VP00	Sjp_0063250	MS3_07393	Sugar transporter
Smp_132730	G4LUC7	Sjp_0108420	MS3_06574	G-protein coupled receptor
Smp_143800	G4VSR1	Sjp_0005570	MS3_01189	Cation transporter
Smp_145900	G4VMQ4	Sjp_0061450	MS3_00627	Dihydroceramide desaturase
Smp_147070	G4VKU1	Sjp_0067720	MS3_02401	Sodium-coupled neutral amino acid transporter
Smp_149450	G4VKS3	Sjp_0026610	MS3_03222	trna-dihydrouridine synthase
Smp_150380	G4VAL8	Sjp_0099280	MS3_07308	Spingomyelin synthetase-related
Smp_155050	G4LY67	Sjp_0002300	MS3_00517	Agrin, Serine protease inhibitor Kazal-type 5-related
Smp_163970	G4VL47	Sjp_0023750	MS3_04617	Carboxypeptidase
Smp_167190	G4V7D5	Sjp_0133070	MS3_05670	Calcium ion binding
Smp_178490	G4LZX3	Sjp_0089300	MS3_08753	Solute carrier family 35

				member d1, UDP-sugar transporter
Smp_180500	G4M0E1	Sjp_0010790	MS3_08237	Phospholipid scramblase-related transfer protein
Smp_199690	G4M1Z7	Sjp_0052970	MS3_09387	G-protein coupled receptor

4.5.4 Protein-protein and protein-chemical interactions

I next examined protein-protein and protein-chemical interactions. Smp_007900.1_mRNA and Smp_050940.1_mRNA, which are 60S ribosomal proteins, effective centre of the network (hub), had most interactions with the 20 antigens (Figure 4.5) and these might have potential as mRNA vaccines. Magnesium Adenosine 5'-triphosphate (MgATP), an adenine nucleotide containing three phosphate groups esterified to the sugar moiety, effective centre of the network (hub), had most interactions with the 20 antigens (Figure 4.6) and might be a potential anti-schistosome drug target.

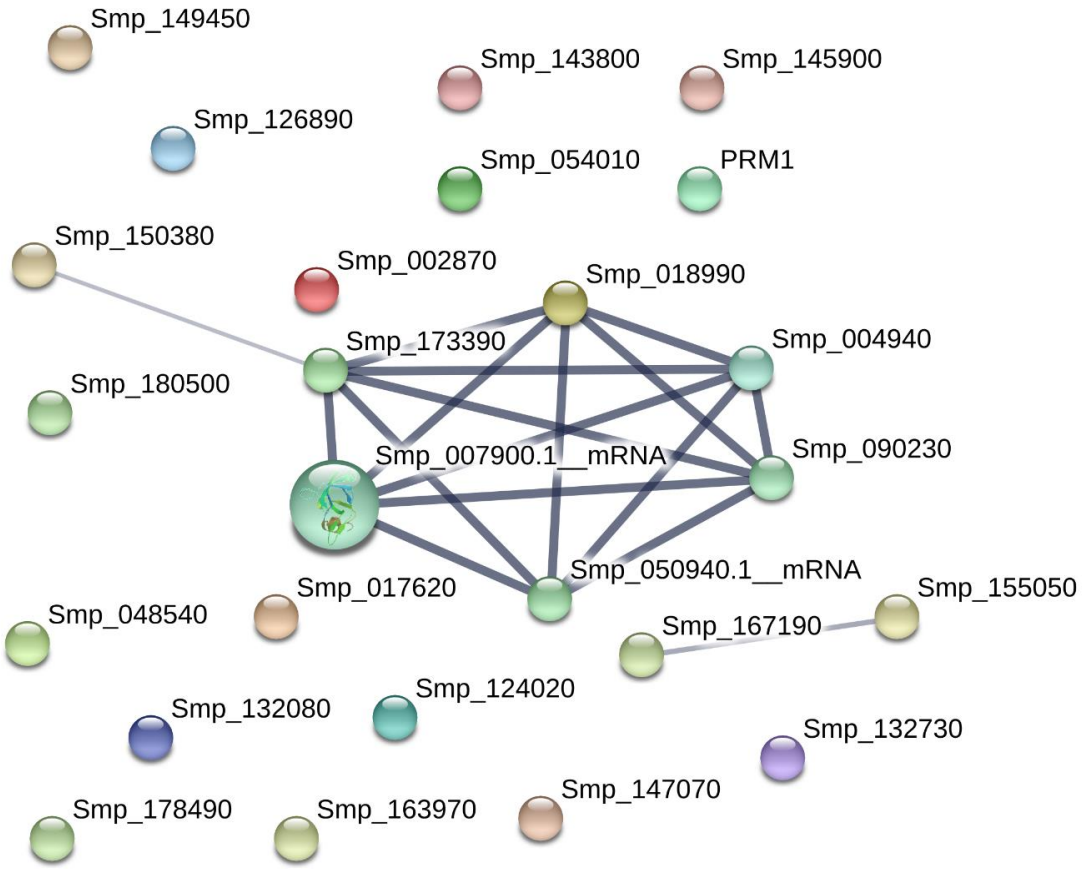


Figure 4.5 Protein-protein interactions for the 20 antigens with other proteins using STRING.

Stronger associations are represented by thicker lines.

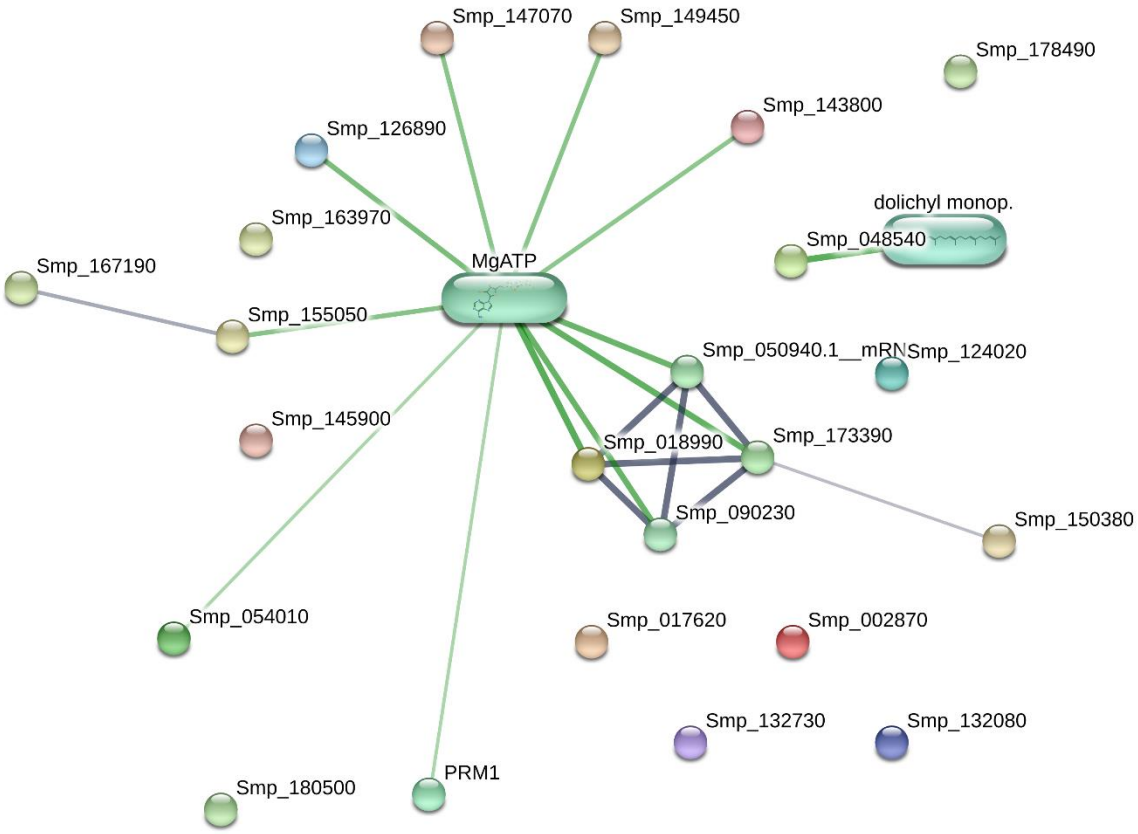


Figure 4.6 Protein-chemical interactions for the 20 antigens using STICH.

Stronger associations are represented by thicker lines. Protein-protein interactions are shown in grey and protein-chemical interactions in green.

4.6 Conclusion

This innovative study provides novel insights into human schistosome genomes and identified gene functions for host-parasite interaction. By using comparative genomics analysis combined with an integrative bioinformatics pipeline I identified putative vaccine antigen candidates and drug targets by assessing their surface and secretory properties, immunogenicity, biological process and molecular function. These novel targets should now be biologically validated by wet laboratory experiments in animals and then clinically. It is particularly noteworthy that many of these molecules have not previously been identified as anti-schistosome intervention targets. The comparative genomics analysis approach for identifying new drug and vaccine candidates represents a valuable resource not only for the *Schistosoma* research community but the protocol I developed can be used as a blueprint for other important parasitic diseases including malaria.

Supporting information

Supplementary Table 4.1 345 core genes of the 3 major schistosome spp. infecting humans which are absent in *S. bovis* and *Schmidtea mediterranea*.

Orthologs were predicted by the RBH method. RBH runs reciprocal BLAST searches against the proteomes of two species (species A vs species and species B vs species A).

The best scoring matches were selected if the High-scoring Segment Pairs (HSPs) had at least 70% identity, 70% alignment length and an E-value of 1e-5.

***S. mansoni* Ortholog**	***S. japonicum* ortholog**	***S. haematobium* ortholog**	**Description**
Smp_165440.1	Sjp_0098730	MS3_06713	Putative netrin receptor unc5
Smp_158970.1	Sjp_0023040	MS3_06013	DEAD box ATP-dependent RNA helicase, putative
Smp_124030.1	Sjp_0000430	MS3_09150	Putative uncharacterized protein
Smp_070680.1	Sjp_0059570	MS3_10545	Putative uncharacterized protein
Smp_203970.1	Sjp_0003090	MS3_08850	Putative uncharacterized protein Smp_203970
Smp_126350.1	Sjp_0018830	MS3_00804	Putative glutamate receptor, NMDA
Smp_038080.1	Sjp_0004060	MS3_03297	Putative importin beta-1
Smp_149640.1	Sjp_0027740	MS3_07138	Putative uncharacterized protein
Smp_149450.1	Sjp_0026610	MS3_03222	Putative trna-dihydrouridine synthase
Smp_147320.1	Sjp_0041890	MS3_06838	Putative camp-dependent protein kinase regulatory chain
Smp_159140.1	Sjp_0034750	MS3_02005	Putative organic anion transporter
Smp_030350.1	Sjp_0012180	MS3_05723	Subfamily S1A unassigned peptidase (S01 family)
Smp_152790.1	Sjp_0053520	MS3_10206	Ras-related GTP binding rag A,B/gtr1
Smp_003250.1	Sjp_0023500	MS3_01004	Putative uncharacterized protein
Smp_069380.1	Sjp_0002050	MS3_03702	Putative histone deacetylase 4, 5
Smp_137580.1	Sjp_0009150	MS3_07623	Helicase, putative
Smp_125590.1	Sjp_0076780	MS3_09275	Putative uncharacterized protein
Smp_179320.1	Sjp_0045200	MS3_08447	Eukaryotic translation initiation factor 2c,putative
Smp_131090.1	Sjp_0115150	MS3_08947	Putative cornichon
Smp_121640.1	Sjp_0113410	MS3_05851	Transcription initiation factor iif (Tfiif),beta subunit-related
Smp_101310.1	Sjp_0068750	MS3_10554	Mizf protein, putative
Smp_046980.1	Sjp_0079920	MS3_02903	Putative uncharacterized protein
Smp_074010.1	Sjp_0080960	MS3_00563	Putative 8-oxoguanine DNA glycosylase
Smp_025130.1	Sjp_0028480	MS3_03667	Putative rna binding motif protein
Smp_159370.1	Sjp_0102090	MS3_01393	Family M13 unassigned peptidase (M13 family)
Smp_033930.1	Sjp_0091400	MS3_10169	Phosphatidylcholine transfer protein, putative
Smp_037900.1	Sjp_0006630	MS3_07488	Family S12 unassigned peptidase (S12 family)
Smp_155330.1	Sjp_0010750	MS3_05152	Serine/threonine kinase
Smp_085680.1	Sjp_0082080	MS3_01929	Guanylate cyclase
Smp_175460.1	Sjp_0003880	MS3_06648	Putative uncharacterized protein
Smp_020220.1	Sjp_0043790	MS3_00034	Putative zeta-coat protein

Smp_145840.1	Sjp_0018790	MS3_01354	Putative wd-repeat protein
Smp_079700.1	Sjp_0074890	MS3_10779	Putative ga binding protein beta chain (Transcription factor e4tf1-47)
Smp_067540.1	Sjp_0064510	MS3_04788	Putative uncharacterized protein
Smp_150470.1	Sjp_0047780	MS3_04903	Putative uncharacterized protein
Smp_044820.1	Sjp_0019750	MS3_06948	Putative uncharacterized protein
Smp_171440.1	Sjp_0115540	MS3_05306	Putative mind bomb
Smp_178850.1	Sjp_0073870	MS3_10823	Poly(A) polymerase, putative
Smp_155610.1	Sjp_0006350	MS3_05713	Calmodulin-5/6/7/8 (CaM-5/6/7/8), putative
Smp_024900.1	Sjp_0034000	MS3_05116	Putative retinoblastoma-like protein
Smp_168130.1	Sjp_0113240	MS3_09956	Phosphatase and actin regulator, putative
Smp_199420.1	Sjp_0088760	MS3_03989	Serine/threonine kinase
Smp_194520.1	Sjp_0006470	MS3_02110	Putative myst histone acetyltransferase
Smp_147920.1	Sjp_0070430	MS3_03684	Ubiquitinyl hydrolase-BAP1 (C12 family)
Smp_136360.1	Sjp_0029860	MS3_10757	Putative dna cross-link repair protein pso2/snm1
Smp_000170.1	Sjp_0004460	MS3_03331	Neurocalcin, putative
Smp_125640.1	Sjp_0066360	MS3_05209	Syntaxin-12, putative
Smp_133040.1	Sjp_0051760	MS3_09170	Putative uncharacterized protein
Smp_012470.1	Sjp_0050140	MS3_11234	Putative 26s protease regulatory subunit
Smp_178490.1	Sjp_0089300	MS3_08753	Solute carrier family 35 member d1, putative
Smp_140530.1	Sjp_0060970	MS3_11377	Putative replication factor C / DNA polymerase III gamma-tau subunit
Smp_150380.1	Sjp_0099280	MS3_07308	Spingomyelin synthetase-related
Smp_017620.1	Sjp_0007190	MS3_05025	Putative uncharacterized protein
Smp_181380.1	Sjp_0101990	MS3_10416	Putative 26s proteasome non-ATPase regulatory subunit
Smp_141860.1	Sjp_0075110	MS3_00212	Putative heat containing protein
Smp_086210.1	Sjp_0002580	MS3_05501	Dihydropteridine reductase
Smp_141470.1	Sjp_0014940	MS3_00828	Putative cytochrome C oxidase assembly protein cox11
Smp_162960.1	Sjp_0037410	MS3_10042	Putative uncharacterized protein
Smp_146830.1	Sjp_0101190	MS3_05591	Putative uncharacterized protein
Smp_006000.1	Sjp_0015700	MS3_01248	Putative eukaryotictranslation initiation factor 3 subunit
Smp_048650.1	Sjp_0114570	MS3_09473	Putative histidine triad (Hit) protein
Smp_148010.1	Sjp_0021030	MS3_06956	Putative snf2 histone linker phd ring helicase
Smp_158920.1	Sjp_0113430	MS3_00867	Putative uracil-DNA glycosylase
Smp_078240.1	Sjp_0110520	MS3_01456	Putative uncharacterized protein
Smp_124310.1	Sjp_0028700	MS3_07274	Putative 5-AMP-activated protein kinase , beta subunit
Smp_002160.1	Sjp_0000610	MS3_04916	Putative uncharacterized protein
Smp_133450.1	Sjp_0001360	MS3_03118	Jnk stimulatory phosphatase-related
Smp_016840.1	Sjp_0040340	MS3_00910	Putative uncharacterized protein
Smp_106930.1	Sjp_0044680	MS3_11411	Heat shock 70 kDa protein homolog
Smp_018640.1	Sjp_0095380	MS3_07841	Putative uncharacterized protein
Smp_026230.1	Sjp_0029800	MS3_07822	Putative uncharacterized protein
Smp_162940.1	Sjp_0055590	MS3_08868	Putative amine oxidase
Smp_136470.1	Sjp_0027350	MS3_08612	Putative rho/rac guanine nucleotide exchange factor

Smp_020200.1	Sjp_0043800	MS3_00007	Putative dead box ATP-dependent RNA helicase
Smp_127180.1	Sjp_0024150	MS3_08448	Putative uncharacterized protein
Smp_166400.1	Sjp_0008030	MS3_09530	Putative dead box ATP-dependent RNA helicase
Smp_054820.1	Sjp_0048190	MS3_03639	Putative uncharacterized protein
Smp_071390.1	Sjp_0065450	MS3_01650	Adenylate kinase
Smp_144950.1	Sjp_0000840	MS3_00778	Putative centrosomal protein of 41 kDa (Cep41 protein) (Testis-specificprotein A14 protein)
Smp_142130.1	Sjp_0134030	MS3_06448	Putative uncharacterized protein
Smp_050130.1	Sjp_0045030	MS3_03863	Putative uncharacterized protein
Smp_132260.1	Sjp_0026310	MS3_00944	Serine/threonine kinase
Smp_094810.1	Sjp_0059010	MS3_04021	Peptidyl-prolyl cis-trans isomerase E
Smp_210570.1	Sjp_0053260	MS3_05559	Mername-AA168 protein (M67 family)
Smp_149000.1	Sjp_0045080	MS3_03862	Protein phosphatase 2C, putative
Smp_139400.1	Sjp_0072070	MS3_01763	Putative tensin
Smp_168670.1	Sjp_0046560	MS3_02488	cGMP-dependent protein kinase,putative
Smp_181350.1	Sjp_0041470	MS3_04867	Huntingtin interacting protein-related
Smp_049890.1	Sjp_0012860	MS3_07669	WD-repeat protein, putative
Smp_153520.1	Sjp_0121100	MS3_08383	Putative uncharacterized protein
Smp_153430.1	Sjp_0085650	MS3_10500	Putative arginyl-tRNA synthetase
Smp_159110.1	Sjp_0067770	MS3_06941	Putative bullous pemphigoid antigen 1, isoform 5 (BPA) (Hemidesmosomal plaque protein) (Dystonia musculorum protein) (Dystonin)
Smp_157820.1	Sjp_0065320	MS3_03496	Putative ataxia telangiectasia mutated (Atm)
Smp_193050.1	Sjp_0029990	MS3_09933	Putative uncharacterized protein
Smp_018890.1	Sjp_0031010	MS3_04778	Phosphoglycerate kinase
Smp_015710.1	Sjp_0076750	MS3_05869	Putative 6-phosphofructo-2-kinase/fructose-2,6-bisphosph at ase
Smp_211290.1	Sjp_0013170	MS3_03109	Putative uncharacterized protein
Smp_130900.1	Sjp_0000570	MS3_01046	Putative alpha catenin
Smp_029310.1	Sjp_0043510	MS3_03741	Phosphatidylinositol transfer protein
Smp_123080.1	Sjp_0044580	MS3_05150	Putative sarcoplasmic calcium-binding protein (SCP)
Smp_133510.1	Sjp_0116270	MS3_00213	Putative aldehyde dehydrogenase
Smp_038300.1	Sjp_0009750	MS3_04095	Putative uncharacterized protein
Smp_132930.1	Sjp_0111900	MS3_06787	Putative uncharacterized protein
Smp_139430.1	Sjp_0078500	MS3_07596	Phosphoinositol 4-phosphate adaptor protein,putative
Smp_196150.1	Sjp_0019390	MS3_10464	Selenoprotein O-like
Smp_105760.1	Sjp_0073210	MS3_00750	Putative innexin
Smp_086460.1	Sjp_0056170	MS3_08022	Tho2 protein, putative
Smp_132550.1	Sjp_0071480	MS3_04780	Putative rhoptry protein
Smp_091770.1	Sjp_0120990	MS3_09612	Protein farnesyltransferase alpha subunit,putative
Smp_210090.1	Sjp_0082760	MS3_02097	Adapter-related protein complex 3, beta subunit
Smp_066250.1	Sjp_0046730	MS3_01279	Putative uncharacterized protein
Smp_155050.1	Sjp_0002300	MS3_00517	Agrin, putative
Smp_152060.1	Sjp_0015940	MS3_05287	Putative uncharacterized protein

Smp_151960.1	Sjp_0085640	MS3_03306	Putative rho gtpase activating protein
Smp_169360.1	Sjp_0045420	MS3_06492	Putative kinesin
Smp_156160.1	Sjp_0067370	MS3_01436	Putative uncharacterized protein
Smp_160700.1	Sjp_0060210	MS3_10140	Putative set domain protein
Smp_035460.1	Sjp_0117330	MS3_07339	Putative nicotinate phosphoribosyltransferase
Smp_079310.1	Sjp_0025150	MS3_03393	Transmembrane protein tmp21-related
Smp_151060.1	Sjp_0018270	MS3_07863	Family S60 non-peptidase homologue (S60 family)
Smp_124500.1	Sjp_0011910	MS3_05903	ADAMTS peptidase (M12 family)
Smp_088660.1	Sjp_0008180	MS3_03445	Putative uncharacterized protein
Smp_024000.1	Sjp_0062010	MS3_07663	Putative vacuolar protein sorting (Vps33)
Smp_159400.1	Sjp_0022690	MS3_01380	Phospholipid transport protein
Smp_092770.1	Sjp_0074500	MS3_10406	Coatomer subunit gamma
Smp_141040.1	Sjp_0133750	MS3_00029	Putative striatin
Smp_136510.1	Sjp_0022020	MS3_05045	Putative uncharacterized protein
Smp_139530.1	Sjp_0071360	MS3_05420	Cellular tumor antigen P53, putative
Smp_150550.1	Sjp_0104790	MS3_02206	Putative titin
Smp_178780.1	Sjp_0106170	MS3_02166	Meso-ectoderm gene expression control protein
Smp_129010.1	Sjp_0057760	MS3_10636	Putative uncharacterized protein
Smp_159890.1	Sjp_0013530	MS3_09839	Metallocarboxypeptidase D peptidase unit 2 (M14 family)
Smp_071840.1	Sjp_0030250	MS3_06507	6-phosphogluconate dehydrogenase, decarboxylating
Smp_158300.1	Sjp_0067840	MS3_01315	Putative uncharacterized protein
Smp_041770.1	Sjp_0000700	MS3_00771	Serine/threonine kinase
Smp_173100.1	Sjp_0024780	MS3_03383	Axon guidance protein
Smp_139070.1	Sjp_0099920	MS3_00607	3bp-1 related rhogap
Smp_132090.1	Sjp_0063260	MS3_07392	Putative wd40 protein
Smp_048540.1	Sjp_0060110	MS3_05665	Putative uncharacterized protein
Smp_147070.1	Sjp_0067720	MS3_02401	Putative amino acid transporter
Smp_131770.1	Sjp_0010590	MS3_03196	DEAD box ATP-dependent RNA helicase, putative
Smp_136750.1	Sjp_0048140	MS3_09438	Putative e3 ubiquitin-protein ligase Bre1 (DBre1)
Smp_055130.1	Sjp_0089630	MS3_09372	Putative zinc finger protein
Smp_167650.1	Sjp_0004330	MS3_01750	Putative uncharacterized protein
Smp_163970.1	Sjp_0023750	MS3_04617	Family S10 non-peptidase homologue (S10 family)
Smp_037200.1	Sjp_0092900	MS3_11285	Putative uncharacterized protein
Smp_012580.1	Sjp_0065880	MS3_10812	Putative guanine-nucleotide-exchange-factor
Smp_134510.1	Sjp_0050230	MS3_01983	Putative uncharacterized protein
Smp_163780.1	Sjp_0051780	MS3_06333	Putative wd-repeat protein
Smp_042030.1	Sjp_0000940	MS3_00766	Putative vesicle transport protein SEC20
Smp_042340.1	Sjp_0003360	MS3_06198	Catenin and plakophilin, putative
Smp_153410.1	Sjp_0029580	MS3_05700	Putative serine/threonine protein phosphatase 2a regulatory subunit A
Smp_154880.1	Sjp_0019670	MS3_10659	Putative tubulin delta chain
Smp_123050.1	Sjp_0044540	MS3_05149	Putative regulator of G protein signaling 17, 19, 20 (Rgs17, 19, 20)
Smp_000510.1	Sjp_0091930	MS3_07584	Gdp-mannose pyrophosphorylase b, isoform 2

Smp_073470.1	Sjp_0002360	MS3_07371	Retinoid-x-receptor (RXR)
Smp_181140.1	Sjp_0087070	MS3_02542	Putative uncharacterized protein
Smp_164340.1	Sjp_0064280	MS3_02998	Afadin (Af-6 protein), putative
Smp_074080.1	Sjp_0035070	MS3_00555	Serine/threonine kinase
Smp_132080.1	Sjp_0063250	MS3_07393	Putative sugar transporter
Smp_120140.1	Sjp_0133110	MS3_08001	Putative uncharacterized protein
Smp_052290.1	Sjp_0089370	MS3_00113	Putative uncharacterized protein
Smp_132500.1	Sjp_0067020	MS3_05998	Putative rab
Smp_054410.1	Sjp_0031570	MS3_09305	Putative adenine phosphoribosyltransferase
Smp_135530.1	Sjp_0071820	MS3_00263	Leishmanolysin-2 (M08 family)
Smp_181240.1	Sjp_0048040	MS3_02816	Glutamyl-tRNA(Gln) amidotransferase subunit B (Mitochondrial and prokaryotic) pet112-related
Smp_179050.1	Sjp_0121200	MS3_09384	Putative kinesin
Smp_142410.1	Sjp_0077070	MS3_03946	Putative uncharacterized protein
Smp_209040.1	Sjp_0053110	MS3_09914	Long-chain-fatty-acid--CoA ligase
Smp_124830.1	Sjp_0069880	MS3_06673	Family C54 unassigned peptidase (C54 family)
Smp_156060.1	Sjp_0053360	MS3_00860	Putative uncharacterized protein
Smp_180500.1	Sjp_0010790	MS3_08237	Phospholipid scramblase-related
Smp_123660.1	Sjp_0049060	MS3_09172	Putative peroxidasin
Smp_084650.1	Sjp_0078000	MS3_09389	Putative uncharacterized protein
Smp_090820.1	Sjp_0027940	MS3_05018	CDP-diacylglycerol--glycerol-3-phosphate 3-phosphatidyltransferase
Smp_016600.1	Sjp_0037260	MS3_00207	Putative solute carrier family 1 (Glial high affinity glutamate transporter)
Smp_048380.1	Sjp_0021110	MS3_06253	Putative uncharacterized protein
Smp_083990.1	Sjp_0048940	MS3_02574	Cationic amino acid transporter, putative
Smp_062560.1	Sjp_0106810	MS3_08312	Putative wnt inhibitor frzb2
Smp_155420.1	Sjp_0118290	MS3_03045	Putative uncharacterized protein
Smp_069600.1	Sjp_0010070	MS3_02274	Putative uncharacterized protein
Smp_061950.1	Sjp_0052690	MS3_09288	Putative u3 small nucleolar ribonucleoprotein
Smp_162450.1	Sjp_0135150	MS3_03335	Cation efflux family protein
Smp_137430.1	Sjp_0025810	MS3_11049	Splicing factor 3B subunit 3, 5'
Smp_198010.1	Sjp_0085760	MS3_02502	Centaurin/arf-related
Smp_136030.1	Sjp_0093940	MS3_07958	Putative anion exchange protein
Smp_174710.1	Sjp_0076400	MS3_05923	Putative spindle assembly checkpoint component MAD1 (Mitotic arrest deficient protein 1)
Smp_194580.1	Sjp_0028000	MS3_05854	Edp1-related
Smp_012030.1	Sjp_0129030	MS3_05333	Putative zinc finger protein
Smp_048240.1	Sjp_0020950	MS3_09733	Putative mannose-1-phosphate guanyltransferase
Smp_001030.1	Sjp_0031410	MS3_01351	5'-amp-activated protein kinase gamma-2 non-catalytic subunit transcript variant 2
Smp_001500.1	Sjp_0071720	MS3_05483	Putative eukaryotic translation initiation factor 4e
Smp_021160.1	Sjp_0043890	MS3_02639	Putative peptidyl-prolyl cis-trans isomerase
Smp_160680.1	Sjp_0053900	MS3_01472	Putative uncharacterized protein

Smp_136320.1	Sjp_0100250	MS3_06428	Putative uncharacterized protein
Smp_084870.1	Sjp_0027180	MS3_06338	Putative uncharacterized protein
Smp_061210.1	Sjp_0002250	MS3_00514	Putative uncharacterized protein
Smp_075470.1	Sjp_0064140	MS3_09296	Cysteine desulfurylase, putative
Smp_019980.1	Sjp_0051690	MS3_08429	Putative vacuole membrane protein
Smp_105100.1	Sjp_0057680	MS3_06302	Ribonuclease, putative
Smp_061940.1	Sjp_0080220	MS3_09286	Putative adenylate kinase 1
Smp_173240.1	Sjp_0012960	MS3_02209	Putative cement protein 3B variant 3
Smp_122340.1	Sjp_0120570	MS3_07583	Kelch-like protein
Smp_036020.1	Sjp_0007850	MS3_07682	Putative uncharacterized protein
Smp_160470.1	Sjp_0044490	MS3_02022	Putative tbc1 domain family member 2 (Prostate antigen recognized and indentified by serex) (Paris-1)
Smp_150220.1	Sjp_0097720	MS3_10047	Putative receptor protein tyrosine phosphatase n, (Ia2)
Smp_171620.1	Sjp_0070480	MS3_05520	S-methyl-5'-thioadenosine phosphorylase
Smp_063110.1	Sjp_0060630	MS3_07840	Putative zinc finger protein
Smp_054010.1	Sjp_0079140	MS3_03571	Putative cationic amino acid transporter
Smp_212140.1	Sjp_0116730	MS3_02920	Vam6/vps39 related
Smp_130890.1	Sjp_0106650	MS3_01037	Putative transient receptor potential cation channel,subfamily m, member
Smp_152500.1	Sjp_0003800	MS3_01807	Putative cyclic-nucleotide-gated cation channel
Smp_203480.1	Sjp_0087200	MS3_09047	Putative uncharacterized protein Smp_203480
Smp_148720.1	Sjp_0010530	MS3_02287	DNA-directed RNA polymerase
Smp_136530.1	Sjp_0022010	MS3_05048	Putative ankyrin repeat-containing
Smp_137370.1	Sjp_0029420	MS3_00329	Serine/threonine kinase
Smp_032780.1	Sjp_0059720	MS3_07190	Putative dolichyl-phosphate beta-glucosyltransferase (dolp-glucosyltransferase)
Smp_163870.1	Sjp_0056320	MS3_03018	Doublesex and mab-3 related transcription factor
Smp_128490.1	Sjp_0104510	MS3_06120	Putative tomosyn
Smp_004360.1	Sjp_0011890	MS3_05904	Putative uncharacterized protein
Smp_176390.1	Sjp_0074340	MS3_04214	Putative cadherin
Smp_016380.1	Sjp_0057120	MS3_11331	Cytohesin-related guanine nucleotide-exchange protein
Smp_124640.1	Sjp_0014090	MS3_09395	Putative dna2/nam7 helicase family member
Smp_073560.1	Sjp_0025130	MS3_07914	G beta-like protein gbl
Smp_168070.1	Sjp_0098810	MS3_00310	Tumor necrosis factor receptor related
Smp_082120.1	Sjp_0082240	MS3_09078	ATP synthase delta chain, mitochondrial,putative
Smp_074710.1	Sjp_0037510	MS3_02415	Putative uncharacterized protein
Smp_020300.1	Sjp_0046610	MS3_00051	Putative dishevelled
Smp_128130.1	Sjp_0097580	MS3_06240	Phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha PI3K
Smp_200110.1	Sjp_0045250	MS3_06589	Putative uncharacterized protein Smp_200110
Smp_036010.1	Sjp_0007840	MS3_09666	Putative uncharacterized protein
Smp_107200.1	Sjp_0081060	MS3_08140	Putative uncharacterized protein
Smp_034550.1	Sjp_0059270	MS3_04427	Putative alpha-actinin
Smp_176800.1	Sjp_0007380	MS3_03148	Putative uncharacterized protein

Smp_151520.1	Sjp_0089670	MS3_02179	Putative dolichyl glycosyltransferase
Smp_123710.1	Sjp_0081470	MS3_05056	Acetyl-CoA carboxylase
Smp_162670.1	Sjp_0065560	MS3_06764	Putative uncharacterized protein
Smp_179560.1	Sjp_0069830	MS3_05089	Multiple inositol polyphosphate phosphatase-related
Smp_091850.1	Sjp_0096790	MS3_11422	Glucosamine-6-phosphate isomerase, putative
Smp_092170.1	Sjp_0074640	MS3_06816	Putative uncharacterized protein
Smp_130040.1	Sjp_0018470	MS3_02806	Pecanex-related protein
Smp_086690.1	Sjp_0071670	MS3_10235	Kinase
Smp_130570.1	Sjp_0027670	MS3_02532	Putative receptor tyrosine phosphatase type r2a
Smp_099930.1	Sjp_0083430	MS3_01056	Putative f-box and wd40 domain protein
Smp_126720.1	Sjp_0058660	MS3_08990	Putative uncharacterized protein
Smp_180740.1	Sjp_0079290	MS3_02664	Putative zinc finger protein
Smp_155590.1	Sjp_0006380	MS3_05716	Putative uncharacterized protein
Smp_171870.1	Sjp_0105540	MS3_06596	Putative sugar transporter
Smp_079390.1	Sjp_0071960	MS3_06639	Protein kinase
Smp_158330.1	Sjp_0037670	MS3_08282	Putative uncharacterized protein
Smp_009540.1	Sjp_0066030	MS3_10330	Putative ring finger-containing
Smp_145900.1	Sjp_0061450	MS3_00627	Dihydroceramide desaturase
Smp_143800.1	Sjp_0005570	MS3_01189	Transporter
Smp_132730.1	Sjp_0108420	MS3_06574	G-protein coupled receptor, putative
Smp_196050.1	Sjp_0114820	MS3_06214	Putative uncharacterized protein Smp_196050
Smp_170010.1	Sjp_0035340	MS3_01416	Putative flagellar radial spoke protein
Smp_174040.1	Sjp_0064790	MS3_08124	Neurotracting/lsamp/neurotrimin/obcam related cell adhesion molecule
Smp_018280.1	Sjp_0108430	MS3_10334	Putative uncharacterized protein
Smp_149020.1	Sjp_0045110	MS3_03870	Lim domain binding protein, putative
Smp_126420.1	Sjp_0070840	MS3_07995	Putative uncharacterized protein
Smp_082370.1	Sjp_0011150	MS3_03842	Putative nadp transhydrogenase
Smp_199690.1	Sjp_0052970	MS3_09387	G-protein coupled receptor,putative
Smp_160940.1	Sjp_0065440	MS3_01661	Putative uncharacterized protein
Smp_179910.1	Sjp_0089480	MS3_00473	Ras-like protein
Smp_002870.1	Sjp_0089930	MS3_07466	Putative uncharacterized protein
Smp_197330.1	Sjp_0081390	MS3_01542	Acetyl-CoA C-acetyltransferase
Smp_055390.1	Sjp_0059520	MS3_05635	Putative dullard protein
Smp_127500.1	Sjp_0070050	MS3_10075	Putative uncharacterized protein
Smp_126800.1	Sjp_0040030	MS3_03936	Putative mkiaa1688 protein
Smp_133320.1	Sjp_0004950	MS3_07197	Ccr4-not transcription complex, putative
Smp_175760.1	Sjp_0002640	MS3_07961	Endosomal trafficking protein, putative
Smp_145670.1	Sjp_0075500	MS3_10068	DNAj-related
Smp_159560.1	Sjp_0011620	MS3_01604	Putative gata binding factor
Smp_170540.1	Sjp_0134980	MS3_04670	Putative uncharacterized protein
Smp_140560.1	Sjp_0030940	MS3_11189	Putative uncharacterized protein
Smp_211090.1	Sjp_0072690	MS3_07969	Putative peptide chain release factor
Smp_169190.1	Sjp_0047050	MS3_05954	Putative tegumental protein\x3b Tegumental allergen-like

			protein
Smp_123780.1	Sjp_0058010	MS3_06562	Putative glypican
Smp_155480.1	Sjp_0105530	MS3_03050	Putative heat shock protein 70 (Hsp70)-interacting protein
Smp_197860.1	Sjp_0066490	MS3_03933	Putative gelsolin
Smp_069170.1	Sjp_0013440	MS3_03548	Putative cation efflux protein/ zinc transporter
Smp_046880.1	Sjp_0052190	MS3_06837	Putative glycosyltransferase
Smp_124570.1	Sjp_0035210	MS3_08722	Leucine zipper protein, putative
Smp_124020.1	Sjp_0094660	MS3_09149	Putative heparan sulfate 6-o-sulfotransferase
Smp_135100.1	Sjp_0028560	MS3_02954	Putative dtdp-glucose 4-6-dehydratase
Smp_093800.1	Sjp_0091070	MS3_03791	Putative uncharacterized protein
Smp_104110.1	Sjp_0040660	MS3_11394	Putative rho GTPase
Smp_211320.1	Sjp_0046960	MS3_05177	Putative soluble guanylate cyclase gcy
Smp_153440.1	Sjp_0029560	MS3_05804	Neuropilin (Nrp) and tolloid (Tll)-like
Smp_167190.1	Sjp_0133070	MS3_05670	Putative uncharacterized protein Smp_167190
Smp_129950.1	Sjp_0066250	MS3_05213	Rna-binding protein 12 (Sh3/ww domain anchor protein in the nucleus) (Swan), putative
Smp_171650.1	Sjp_0059410	MS3_04240	Eps-15-related
Smp_158670.1	Sjp_0031200	MS3_05239	Putative uncharacterized protein
Smp_168220.1	Sjp_0027270	MS3_06346	Putative uncharacterized protein
Smp_036590.1	Sjp_0047920	MS3_00233	Putative polypyrimidine tract binding protein
Smp_166420.1	Sjp_0008060	MS3_06923	Putative uncharacterized protein
Smp_038640.1	Sjp_0020590	MS3_04345	Cactin-related
Smp_146400.1	Sjp_0024890	MS3_03372	Syntaxin binding protein-1,2,3, putative
Smp_141660.1	Sjp_0006720	MS3_06102	Putative high voltage-activated calcium channel beta subunit 2
Smp_144130.1	Sjp_0037860	MS3_01201	Septate junction protein
Smp_161510.1	Sjp_0008800	MS3_08477	Putative uncharacterized protein
Smp_024290.1	Sjp_0088830	MS3_06556	Putative map kinase kinase protein DdMEK1
Smp_139020.1	Sjp_0029160	MS3_00600	Protein kinase
Smp_070380.1	Sjp_0055710	MS3_10499	Putative uncharacterized protein
Smp_148530.1	Sjp_0107970	MS3_03420	Putative heat shock protein hsp16
Smp_129260.1	Sjp_0086100	MS3_11002	Poly
Smp_049880.1	Sjp_0100280	MS3_07670	Nuclear pore glycoprotein P62, putative
Smp_087560.1	Sjp_0032790	MS3_04593	Putative uncharacterized protein
Smp_053550.1	Sjp_0052010	MS3_06215	Putative uncharacterized protein
Smp_170530.1	Sjp_0084910	MS3_04667	Upstream transcription factor 1, usf1, putative
Smp_071950.1	Sjp_0019810	MS3_08816	Tetratricopeptide repeat protein, tpr, putative
Smp_021170.1	Sjp_0043880	MS3_02636	VPS13C protein, putative
Smp_168090.1	Sjp_0100660	MS3_10683	Putative uncharacterized protein
Smp_087620.1	Sjp_0054960	MS3_04595	Putative programmed cell death 6-interacting protein
Smp_012350.1	Sjp_0070310	MS3_11409	Venom allergen-like (VAL) 11 protein
Smp_125990.1	Sjp_0085480	MS3_09587	Putative uncharacterized protein
Smp_142890.1	Sjp_0069480	MS3_04099	Putative hect E3 ubiquitin ligase
Smp_090070.1	Sjp_0085730	MS3_02507	WD-repeat protein, putative
Smp_082800.1	Sjp_0030400	MS3_08257	Sly1-related

Smp_134230.1	Sjp_0093830	MS3_09190	Putative geranylgeranyl pyrophosphate synthase
Smp_131970.1	Sjp_0063890	MS3_02420	Putative uncharacterized protein
Smp_093920.1	Sjp_0129510	MS3_07544	Putative uncharacterized protein
Smp_154690.1	Sjp_0048310	MS3_04706	Ribosomal protein related
Smp_030710.1	Sjp_0055890	MS3_06743	Serine/threonine-protein phosphatase
Smp_134370.1	Sjp_0015320	MS3_07853	Putative rna 3' terminal phosphate cyclase
Smp_199400.1	Sjp_0069800	MS3_05091	Myosin regulatory light chain 2 smooth muscle,putative
Smp_016410.1	Sjp_0088510	MS3_08949	Putative ran-binding protein
Smp_173060.1	Sjp_0018530	MS3_07261	Pak-interacting exchange factor, beta-pix/cool-1, putative
Smp_036500.1	Sjp_0106460	MS3_00211	CAR\x3b Putative nuclear receptor nhr-48
Smp_098560.2	Sjp_0052120	MS3_06205	Putative uncharacterized protein
Smp_168590.1	Sjp_0054800	MS3_02999	Macrophage scavenger receptor-related
Smp_150210.1	Sjp_0026510	MS3_09307	Putative autocrine motility factor receptor,amfr
Smp_138670.1	Sjp_0038050	MS3_09404	39S ribosomal protein L19, mitochondrial, putative
Smp_123470.1	Sjp_0100570	MS3_00446	Nalp (Nacht, leucine rich repeat and pyrin domain containing)-related
Smp_175720.1	Sjp_0130370	MS3_03583	Tripartite motif protein trim9, putative
Smp_040800.1	Sjp_0039430	MS3_08104	Putative glycyl-tRNA synthetase
Smp_166960.1	Sjp_0042790	MS3_06050	Putative ubiquitin-protein ligase BRE1
Smp_176420.1	Sjp_0030500	MS3_08694	Ras guanine nucleotide exchange factor ,putative
Smp_088950.1	Sjp_0060680	MS3_10049	Hypoxia upregulated 1 (Hyou1)-related
Smp_160710.1	Sjp_0054490	MS3_07504	Putative upstream stimulatory factor
Smp_018990.1	Sjp_0075950	MS3_04781	Putative 60s ribosomal protein L9
Smp_128860.1	Sjp_0050160	MS3_10980	Lysyl oxidase-like
Smp_063000.1	Sjp_0012640	MS3_01088	Smdr1
Smp_157350.1	Sjp_0049520	MS3_01308	Protein kinase
Smp_147240.1	Sjp_0031690	MS3_02755	Putative wd40 protein

Supplementary Table 4.2 135 proteins were predicted as surface or secretory proteins by SchistoProt and SchistoTarget.

Protein ID	**Surface Prediction**	**Secretory Prediction**
Smp_000170.1	Surface Protein	Secretory Protein
Smp_001030.1	Surface Protein	Secretory Protein
Smp_003250.1	Surface Protein	Secretory Protein
Smp_004360.1	Surface Protein	Secretory Protein
Smp_006000.1	Surface Protein	Secretory Protein
Smp_015710.1	Surface Protein	Secretory Protein
Smp_016600.1	Surface Protein	Secretory Protein
Smp_018890.1	Surface Protein	Secretory Protein

Smp_018990.1	Surface Protein	Secretory Protein
Smp_019980.1	Surface Protein	Secretory Protein
Smp_021160.1	Surface Protein	Secretory Protein
Smp_026230.1	Surface Protein	Secretory Protein
Smp_029310.1	Surface Protein	Secretory Protein
Smp_033930.1	Surface Protein	Secretory Protein
Smp_034550.1	Surface Protein	Secretory Protein
Smp_048380.1	Surface Protein	Secretory Protein
Smp_048540.1	Surface Protein	Secretory Protein
Smp_061210.1	Surface Protein	Secretory Protein
Smp_061940.1	Surface Protein	Secretory Protein
Smp_067540.1	Surface Protein	Secretory Protein
Smp_070380.1	Surface Protein	Secretory Protein
Smp_071390.1	Surface Protein	Secretory Protein
Smp_071840.1	Surface Protein	Secretory Protein
Smp_079310.1	Surface Protein	Secretory Protein
Smp_084650.1	Surface Protein	Secretory Protein
Smp_087560.1	Surface Protein	Secretory Protein
Smp_094810.1	Surface Protein	Secretory Protein
Smp_104110.1	Surface Protein	Secretory Protein
Smp_105100.1	Surface Protein	Secretory Protein
Smp_105760.1	Surface Protein	Secretory Protein
Smp_106930.1	Surface Protein	Secretory Protein
Smp_123080.1	Surface Protein	Secretory Protein
Smp_123470.1	Surface Protein	Secretory Protein
Smp_124020.1	Surface Protein	Secretory Protein
Smp_124570.1	Surface Protein	Secretory Protein
Smp_126720.1	Surface Protein	Secretory Protein
Smp_135100.1	Surface Protein	Secretory Protein
Smp_140560.1	Surface Protein	Secretory Protein
Smp_141470.1	Surface Protein	Secretory Protein
Smp_145900.1	Surface Protein	Secretory Protein
Smp_150380.1	Surface Protein	Secretory Protein
Smp_151520.1	Surface Protein	Secretory Protein
Smp_151960.1	Surface Protein	Secretory Protein
Smp_152790.1	Surface Protein	Secretory Protein

Smp_169190.1	Surface Protein	Secretory Protein
Smp_178490.1	Surface Protein	Secretory Protein
Smp_179910.1	Surface Protein	Secretory Protein
Smp_180500.1	Surface Protein	Secretory Protein
Smp_193050.1	Surface Protein	Secretory Protein
Smp_196050.1	Surface Protein	Secretory Protein
Smp_199400.1	Surface Protein	Secretory Protein
Smp_199420.1	Surface Protein	Secretory Protein
Smp_200110.1	Surface Protein	Secretory Protein
Smp_203970.1	Surface Protein	Secretory Protein
Smp_001500.1	Surface Protein	Non-Secretory Protein
Smp_017620.1	Surface Protein	Non-Secretory Protein
Smp_020220.1	Surface Protein	Non-Secretory Protein
Smp_030710.1	Surface Protein	Non-Secretory Protein
Smp_038300.1	Surface Protein	Non-Secretory Protein
Smp_044820.1	Surface Protein	Non-Secretory Protein
Smp_046980.1	Surface Protein	Non-Secretory Protein
Smp_048650.1	Surface Protein	Non-Secretory Protein
Smp_054410.1	Surface Protein	Non-Secretory Protein
Smp_073560.1	Surface Protein	Non-Secretory Protein
Smp_091770.1	Surface Protein	Non-Secretory Protein
Smp_091850.1	Surface Protein	Non-Secretory Protein
Smp_098560.2	Surface Protein	Non-Secretory Protein
Smp_107200.1	Surface Protein	Non-Secretory Protein
Smp_120140.1	Surface Protein	Non-Secretory Protein
Smp_127500.1	Surface Protein	Non-Secretory Protein
Smp_131970.1	Surface Protein	Non-Secretory Protein
Smp_134510.1	Surface Protein	Non-Secretory Protein
Smp_136320.1	Surface Protein	Non-Secretory Protein
Smp_149450.1	Surface Protein	Non-Secretory Protein
Smp_158330.1	Surface Protein	Non-Secretory Protein
Smp_158970.1	Surface Protein	Non-Secretory Protein
Smp_162960.1	Surface Protein	Non-Secretory Protein
Smp_163870.1	Surface Protein	Non-Secretory Protein
Smp_166400.1	Surface Protein	Non-Secretory Protein
Smp_168220.1	Surface Protein	Non-Secretory Protein

Smp_170010.1	Surface Protein	Non-Secretory Protein
Smp_171620.1	Surface Protein	Non-Secretory Protein
Smp_209040.1	Surface Protein	Non-Secretory Protein
Smp_000510.1	Non-Surface Protein	Secretory Protein
Smp_002870.1	Non-Surface Protein	Secretory Protein
Smp_012470.1	Non-Surface Protein	Secretory Protein
Smp_018640.1	Non-Surface Protein	Secretory Protein
Smp_030350.1	Non-Surface Protein	Secretory Protein
Smp_035460.1	Non-Surface Protein	Secretory Protein
Smp_036010.1	Non-Surface Protein	Secretory Protein
Smp_037900.1	Non-Surface Protein	Secretory Protein
Smp_048240.1	Non-Surface Protein	Secretory Protein
Smp_050130.1	Non-Surface Protein	Secretory Protein
Smp_054010.1	Non-Surface Protein	Secretory Protein
Smp_061950.1	Non-Surface Protein	Secretory Protein
Smp_069170.1	Non-Surface Protein	Secretory Protein
Smp_074010.1	Non-Surface Protein	Secretory Protein
Smp_075470.1	Non-Surface Protein	Secretory Protein
Smp_083990.1	Non-Surface Protein	Secretory Protein
Smp_085680.1	Non-Surface Protein	Secretory Protein
Smp_086210.1	Non-Surface Protein	Secretory Protein
Smp_088950.1	Non-Surface Protein	Secretory Protein
Smp_093800.1	Non-Surface Protein	Secretory Protein
Smp_125640.1	Non-Surface Protein	Secretory Protein
Smp_125990.1	Non-Surface Protein	Secretory Protein
Smp_126350.1	Non-Surface Protein	Secretory Protein
Smp_128860.1	Non-Surface Protein	Secretory Protein
Smp_130040.1	Non-Surface Protein	Secretory Protein
Smp_132080.1	Non-Surface Protein	Secretory Protein
Smp_132500.1	Non-Surface Protein	Secretory Protein
Smp_132730.1	Non-Surface Protein	Secretory Protein
Smp_132930.1	Non-Surface Protein	Secretory Protein
Smp_136360.1	Non-Surface Protein	Secretory Protein
Smp_138670.1	Non-Surface Protein	Secretory Protein
Smp_143800.1	Non-Surface Protein	Secretory Protein
Smp_144130.1	Non-Surface Protein	Secretory Protein

Smp_146400.1	Non-Surface Protein	Secretory Protein
Smp_146830.1	Non-Surface Protein	Secretory Protein
Smp_147070.1	Non-Surface Protein	Secretory Protein
Smp_150210.1	Non-Surface Protein	Secretory Protein
Smp_155050.1	Non-Surface Protein	Secretory Protein
Smp_155420.1	Non-Surface Protein	Secretory Protein
Smp_155610.1	Non-Surface Protein	Secretory Protein
Smp_159370.1	Non-Surface Protein	Secretory Protein
Smp_159400.1	Non-Surface Protein	Secretory Protein
Smp_159890.1	Non-Surface Protein	Secretory Protein
Smp_162450.1	Non-Surface Protein	Secretory Protein
Smp_163970.1	Non-Surface Protein	Secretory Protein
Smp_167190.1	Non-Surface Protein	Secretory Protein
Smp_173100.1	Non-Surface Protein	Secretory Protein
Smp_178780.1	Non-Surface Protein	Secretory Protein
Smp_179560.1	Non-Surface Protein	Secretory Protein
Smp_199690.1	Non-Surface Protein	Secretory Protein
Smp_211090.1	Non-Surface Protein	Secretory Protein
Smp_211320.1	Non-Surface Protein	Secretory Protein

Supplementary Table 4.3 SchistoTarget predicted 45 proteins have immunoreactivity among the 135 proteins.

Immunoreactive proteins		
Smp_006000.1	Smp_017620.1	Smp_132080.1
Smp_018990.1	Smp_048650.1	Smp_132730.1
Smp_048380.1	Smp_120140.1	Smp_132930.1
Smp_048540.1	Smp_127500.1	Smp_136360.1
Smp_070380.1	Smp_149450.1	Smp_143800.1
Smp_105760.1	Smp_162960.1	Smp_147070.1
Smp_124020.1	Smp_163870.1	Smp_155050.1
Smp_140560.1	Smp_168220.1	Smp_159400.1
Smp_145900.1	Smp_002870.1	Smp_163970.1
Smp_150380.1	Smp_030350.1	Smp_167190.1
Smp_151520.1	Smp_036010.1	Smp_173100.1
Smp_178490.1	Smp_054010.1	Smp_178780.1
Smp_180500.1	Smp_083990.1	Smp_179560.1

Smp_196050.1	Smp_125990.1	Smp_199690.1
Smp_200110.1	Smp_130040.1	Smp_211320.1

Supplementary Table 4.4 20 proteins were selected as potential vaccine targets using GO annotation.

Proteins were selected on the basis of biological processes and molecular functions which are important for host-parasite interactions such as catalytic activity, transmembrane transporter activity, lipid transporter activity, serine-type peptidase activity, serine protease inhibitory actvity, G-protein coupled receptor and oxidoreductase activity.

Gene ID	PANTHER GO-Slim Molecular Function	PANTHER GO-Slim Biological Process
Smp_002870	protein binding(GO:0005515); small GTPase regulator activity(GO:0005083)	cellular process(GO:0009987)
Smp_017620		
Smp_018990	RNA binding(GO:0003723); structural constituent of ribosome(GO:0003735)	biosynthetic process(GO:0009058); cellular process(GO:0009987); translation(GO:0006412)
Smp_048540		
Smp_054010	amino acid transmembrane transporter activity(GO:0015171); transmembrane transporter activity(GO:0022857)	amino acid transport(GO:0006865); anion transport(GO:0006820); cellular process(GO:0009987)
Smp_083990	amino acid transmembrane transporter activity(GO:0015171); transmembrane transporter activity(GO:0022857)	amino acid transport(GO:0006865); anion transport(GO:0006820); cellular process(GO:0009987)
Smp_124020	transferase activity(GO:0016740)	biosynthetic process(GO:0009058); cellular process(GO:0009987); protein metabolic process(GO:0019538); sulfur compound metabolic process(GO:0006790)
Smp_132080	transmembrane transporter activity(GO:0022857)	cellular process(GO:0009987)
Smp_132730		
Smp_143800	transporter activity(GO:0005215)	
Smp_145900	oxidoreductase activity(GO:0016491)	biosynthetic process(GO:0009058); cellular process(GO:0009987); lipid metabolic process(GO:0006629); nitrogen compound metabolic process(GO:0006807)
Smp_147070	amino acid transmembrane transporter activity(GO:0015171); transmembrane transporter activity(GO:0022857)	amino acid transport(GO:0006865); anion transport(GO:0006820); cellular process(GO:0009987)
Smp_149450		
Smp_150380	catalytic activity(GO:0003824)	lipid metabolic process(GO:0006629)
Smp_155050		
Smp_163970	serine-type peptidase activity(GO:0008236)	catabolic process(GO:0009056); cellular process(GO:0009987); proteolysis(GO:0006508)
Smp_167190	calcium ion binding(GO:0005509); calcium-	cellular process(GO:0009987)

	dependent phospholipid binding(GO:0005544); calmodulin binding(GO:0005516); extracellular matrix structural constituent(GO:0005201); receptor binding(GO:0005102)	
Smp_178490	transmembrane transporter activity(GO:0022857)	cellular process(GO:0009987); nucleobase-containing compound transport(GO:0015931)
Smp_180500	lipid transporter activity(GO:0005319)	anion transport(GO:0006820); biological regulation(GO:0065007); cellular component organization(GO:0016043); cellular process(GO:0009987)
Smp_199690	G-protein coupled receptor activity(GO:0004930); binding(GO:0005488); signal transducer activity(GO:0004871)	regulation of biological process(GO:0050789); response to endogenous stimulus(GO:0009719)

Chapter 5 General discussion and conclusion

This chapter discussed and summarised the overall PhD research works and findings. This PhD thesis covers three individual projects that are building up on each other to achieve the three aims of the research. Chapter 1 describes the theoretical background that is required to understand the thesis topics and gives a good and comprehensive overview of schistosomiasis, its causes, and consequences. Chapters 2 to chapter 4 describe each one of the three aims of the PhD research. Chapter 2 fulfils the aim-1 which is the development of a method for the identification of schistosome-specific surface proteins and secreted peptides. Aim 2 is achieved in chapter 3, the development of a method for the identification of Schistosoma immunoreactive proteins. The final aim of the PhD research is mentioned in chapter 4, the application of the developed methods in an integrative bioinformatics pipeline to identify putative vaccine targets against schistosomiasis.

Schistosomiasis is a parasitic disease caused by parasitic *Schistosoma*. This disease is the second most devastating parasitic disease after malaria worms and more than 200 million people are infected worldwide. Despite of deadly effect on mass population, schistosomiasis is considered one of the Neglected Tropical Diseases (NTDs). No vaccines are available and treatment relies mainly on one drug, praziquantel. This drug effectively disrupts the tegument of adult worms, but not juvenile parasites and even mass treatment does not prevent reinfection. Vaccines that induce long-term immunity represent an essential component for the future control of schistosomiasis with the final goal of complete elimination. Driven by the need to improve disease treatment and prevention, the genomes of several *Schistosoma* species have recently become publicly available. Moreover, several whole-genome sequencing projects of additional *Schistosoma* species will soon be completed.

Surface-associated proteins and secreted peptides play a key role in parasite physiology and pathogenesis and are the major targets for vaccine development. The laborious task of identifying these important classes of proteins can be highly accelerated using computational tools that evaluate the specific sequence properties of surface proteins and secreted peptides. However, currently available methods for this task have their limitations as they show only modest prediction accuracy for *Schistosoma* species. SchistoProt, a machine-learning classifier for the identification of *Schistosoma* surface proteins and

secreted peptides, have been developed. SchistoProt provides a user-friendly web-interface and achieves a superior detection accuracy compared to other existing tools. Results are presented as interactive tables, charts and figures. This project results demonstrate that a genus-specific classifier can excel in the detection of surface proteins and secreted peptides compared with general tools for this task. As such, SchistoProt assists in studying the molecular mechanisms of host infection, analysing anti-schistosome protective immunity and the rapid prioritization of candidate vaccine targets.

Schistosome protein microarrays allow to compare antibody signatures in different disease pathologies as the pilot *Schistosoma* immunomics study. SchistoTarget enables us to identify *Schistosoma* proteins immunoreactivity using the small size of protein microarray data by machine learning based approach. The prediction accuracy of SchistoTarget can be further improved by training on future protein microarrays data for antibody responses, if available.

The innovative integrative approach developed with the comparative studies of three human-infecting schistosomes, animal-infecting *S. bovis* and non-parasitic free-living flatworm *Schmidtea mediterranea* reveals the interesting candidate genes for vaccine targets. These genes are further filtered using the tools developed in this PhD project, SchistoProt and SchistoTarget, to select only surface or secretory proteins. Gene Ontology and Swiss-Prot annotations provide the useful information to select potential antigens as drug and vaccine targets. Further, applying STRING and STICH, protein-protein and protein-chemical interactions are explored which provided putative vaccines and drug targets.

The results of this PhD project are expected to significance advances in schsitosomiasis vaccinology. The selected 20 antigens as putative vaccine targets against schistosomiasis should now be biologically validated by wet laboratory experiments in animals and then clinically. The protocol developed in this PhD project can be used as a blueprint for other parasitic diseases such as malaria.

References

1 Lawton, S. P., Hirai, H., Ironside, J. E., Johnston, D. A. & Rollinson, D. Genomes and geography: genomic insights into the evolution and phylogeography of the genus Schistosoma. *Parasites & Vectors* **4**, 131, doi:10.1186/1756-3305-4-131 (2011).

2 McWilliam, H. E. G., Driguez, P., Piedrafita, D., McManus, D. P. & Meeusen, E. N. T. Novel immunomic technologies for schistosome vaccine development. *Parasite Immunology* **34**, 276-284, doi:10.1111/j.1365-3024.2011.01330.x (2012).

3 van der Werf, M. J. *et al.* Quantification of clinical morbidity associated with schistosome infection in sub-Saharan Africa. *Acta Tropica* **86**, 125-139, doi:10.1016/S0001-706X(03)00029-9 (2003).

4 Webster, B. L., Southgate, V. R. & Littlewood, D. T. J. A revision of the interrelationships of Schistosoma including the recently described Schistosoma guineensis. *International Journal for Parasitology* **36**, 947-955, doi:10.1016/j.ijpara.2006.03.005 (2006).

5 Cantacessi, C. *et al.* Bioinformatics meets parasitology. *Parasite Immunology* **34**, 265-275, doi:10.1111/j.1365-3024.2011.01304.x (2012).

6 Standley, C. J., Dobson, A. P. & Stothard, J. R. *Out of Animals and Back Again: Schistosomiasis as a Zoonosis in Africa, Schistosomiasis*. (InTech, 2012).

7 Skelly, P. J. & Alan Wilson, R. Making Sense of the Schistosome Surface. *Advances in Parasitology* **63**, 185-284, doi:10.1016/S0065-308X(06)63003-0 (2006).

8 Ross , A. G. P. *et al.* Schistosomiasis. *New England Journal of Medicine* **346**, 1212-1220, doi:10.1056/NEJMra012396 (2002).

9 Gryseels, B., Polman, K., Clerinx, J. & Kestens, L. Human schistosomiasis. *The Lancet* **368**, 1106-1118, doi:10.1016/S0140-6736(06)69440-3 (2006).

10 Wang, L., Li, Y.-L., Fishelson, Z., Kusel, J. R. & Ruppel, A. Schistosoma japonicum migration through mouse skin compared histologically and immunologically with S. mansoni. *Parasitology Research* **95**, 218-223, doi:10.1007/s00436-004-1284-4 (2005).

11 Wang, W., Wang, L. & Liang, Y.-S. Susceptibility or resistance of praziquantel in human schistosomiasis: a review. *Parasitology Research* **111**, 1871-1877, doi:10.1007/s00436-012-3151-z (2012).

12 Gray, D. J. *et al.* Schistosomiasis elimination: lessons from the past guide the future. *The Lancet Infectious Diseases* **10**, 733-736, doi:10.1016/S1473-3099(10)70099-2.

13 Doenhoff, M. J., Kusel, J. R., Coles, G. C. & Cioli, D. Resistance of Schistosoma mansoni to praziquantel: is there a problem? *Transactions of the Royal Society of Tropical Medicine and Hygiene* **96**, 465-469, doi:10.1016/S0035-9203(02)90405-0 (2002).

14 Liu, R., Dong, H.-F., Guo, Y., Zhao, Q.-P. & Jiang, M.-S. Efficacy of praziquantel and artemisinin derivatives for the treatment and prevention of human schistosomiasis: a systematic review and meta-analysis. *Parasites & Vectors* **4**, 201, doi:10.1186/1756-3305-4-201 (2011).

15 Gaze, S. *et al.* An Immunomics Approach to Schistosome Antigen Discovery: Antibody Signatures of Naturally Resistant and Chronically Infected Individuals from Endemic Areas. *PLOS Pathogens* **10**, e1004033, doi:10.1371/journal.ppat.1004033 (2014).

16 Riveau, G. *et al.* Safety and Immunogenicity of rSh28GST Antigen in Humans: Phase 1 Randomized Clinical Study of a Vaccine Candidate against Urinary Schistosomiasis. *PLOS Neglected Tropical Diseases* **6**, e1704, doi:10.1371/journal.pntd.0001704 (2012).

17 Tran, M. H. *et al.* Tetraspanins on the surface of Schistosoma mansoni are protective antigens against schistosomiasis. *Nat Med* **12**, 835-840, doi:10.1038/nm1430 (2006).

18 Tendler, M. & Simpson, A. J. G. The biotechnology-value chain: Development of Sm14 as a schistosomiasis vaccine. *Acta Tropica* **108**, 263-266, doi:10.1016/j.actatropica.2008.09.002 (2008).

19 Hotez, P. J., Bethony, J. M., Diemert, D. J., Pearson, M. & Loukas, A. Developing vaccines to combat hookworm infection and intestinal schistosomiasis. *Nat Rev Micro* **8**, 814-826 (2010).

20 Berriman, M. *et al.* The genome of the blood fluke Schistosoma mansoni. *Nature* **460**, 352-358, doi:10.1038/nature08160 (2009).

21 Zhou, Y. *et al.* The Schistosoma japonicum genome reveals features of host-parasite interplay. *Nature* **460**, 345-351, doi:10.1038/nature08140 (2009).

22 Young, N. D. *et al.* Whole-genome sequence of Schistosoma haematobium. *Nat Genet* **44**, 221-225, doi:10.1038/ng.1065 (2012).

23 Braschi, S. & Wilson, R. A. Proteins Exposed at the Adult Schistosome Surface Revealed by Biotinylation. *Molecular & Cellular Proteomics* **5**, 347-356, doi:10.1074/mcp.M500287-MCP200 (2006).

24 Mulvenna, J. *et al.* Exposed proteins of the Schistosoma japonicum tegument. *International Journal for Parasitology* **40**, 543-554, doi:10.1016/j.ijpara.2009.10.002 (2010).

25 Jenkins, S. J., Hewitson, J. P., Jenkins, G. R. & Mountford, A. P. Modulation of the host's immune response by schistosome larvae. *Parasite Immunology* **27**, 385-393, doi:10.1111/j.1365-3024.2005.00789.x (2005).

26 Castro-Borges, W. *et al.* Abundance of tegument surface proteins in the human blood fluke Schistosoma mansoni determined by QconCAT proteomics. *Journal of Proteomics* **74**, 1519-1533, doi:10.1016/j.jprot.2011.06.011 (2011).

27 Liao, Q. *et al.* Identifying Schistosoma japonicum Excretory/Secretory Proteins and Their Interactions with Host Immune System. *PLOS ONE* **6**, e23786, doi:10.1371/journal.pone.0023786 (2011).

28 Liu, F. *et al.* New Perspectives on Host-Parasite Interplay by Comparative Transcriptomic and Proteomic Analyses of Schistosoma japonicum. *PLOS Pathogens* **2**, e29, doi:10.1371/journal.ppat.0020029 (2006).

29 Harris, A. R. C., Russell, R. J. & Charters, A. D. A review of schistosomiasis in immigrants in Western Australia, demonstrating the unusual longevity of Schistosoma mansoni. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **78**, 385-388, doi:10.1016/0035-9203(84)90129-9 (1984).

30 Han, Z.-G., Brindley, P. J., Wang, S.-Y. & Chen, Z. Schistosoma Genomics: New Perspectives on Schistosome Biology and Host-Parasite Interaction. *Annual Review of Genomics and Human Genetics* **10**, 211-240, doi:10.1146/annurev-genom-082908-150036 (2009).

31 Fonseca, C. T. *et al.* Schistosoma Tegument Proteins in Vaccine and Diagnosis Development: An Update. *Journal of Parasitology Research* **2012**, 8, doi:10.1155/2012/541268 (2012).

32 van Balkom, B. W. M. *et al.* Mass Spectrometric Analysis of the Schistosoma mansoni Tegumental Sub-proteome. *Journal of Proteome Research* **4**, 958-966, doi:10.1021/pr050036w (2005).

33 Teixeira de Melo, T. *et al.* Immunization with newly transformed Schistosoma mansoni schistosomula tegument elicits tegument damage, reduction in egg and parasite burden. *Parasite Immunology* **32**, 749-759, doi:10.1111/j.1365-3024.2010.01244.x (2010).

34 Ridi, R. E. & Tallima, H. Schistosoma mansoni ex vivo lung-stage larvae excretory–secretory antigens as vaccine candidates against schistosomiasis. *Vaccine* **27**, 666-673, doi:10.1016/j.vaccine.2008.11.039 (2009).

35 McManus, D. P. & Loukas, A. Current status of vaccines for schistosomiasis. *Clinical microbiology reviews* **21**, 225-242, doi:10.1128/cmr.00046-07 (2008).

36 de Assis, R. R. *et al.* A next-generation proteome array for Schistosoma mansoni. *International Journal for Parasitology* **46**, 411-415, doi:10.1016/j.ijpara.2016.04.001 (2016).

37 Braschi, S., Curwen, R. S., Ashton, P. D., Verjovski-Almeida, S. & Wilson, A. The tegument surface membranes of the human blood parasite Schistosoma mansoni: A proteomic analysis after differential extraction. *PROTEOMICS* **6**, 1471-1482, doi:10.1002/pmic.200500368 (2006).

38 Cass, C. L. *et al.* Proteomic analysis of Schistosoma mansoni egg secretions. *Molecular and Biochemical Parasitology* **155**, 84-93, doi:10.1016/j.molbiopara.2007.06.002 (2007).

39 Sotillo, J., Pearson, M., Becker, L., Mulvenna, J. & Loukas, A. A quantitative proteomic analysis of the tegumental proteins from Schistosoma mansoni schistosomula reveals novel potential therapeutic targets. *International Journal for Parasitology* **45**, 505-516, doi:10.1016/j.ijpara.2015.03.004 (2015).

40 Van Hellemond, J. J. *et al.* Functions of the tegument of schistosomes: Clues from the proteome and lipidome. *International Journal for Parasitology* **36**, 691-699, doi:10.1016/j.ijpara.2006.01.007 (2006).

41 Walker, A. J. Insights into the functional biology of schistosomes. *Parasites & Vectors* **4**, 203, doi:10.1186/1756-3305-4-203 (2011).

42 Castro-Borges, W., Dowle, A., Curwen, R. S., Thomas-Oates, J. & Wilson, R. A. Enzymatic Shaving of the Tegument Surface of Live Schistosomes for Proteomic

Analysis: A Rational Approach to Select Vaccine Candidates. *PLOS Neglected Tropical Diseases* **5**, e993, doi:10.1371/journal.pntd.0000993 (2011).

43 Liu, F. *et al.* Excretory/Secretory Proteome of the Adult Developmental Stage of Human Blood Fluke, Schistosoma japonicum. *Molecular & Cellular Proteomics* **8**, 1236-1251, doi:10.1074/mcp.M800538-MCP200 (2009).

44 Dvořák, J. *et al.* Excretion/secretion products from Schistosoma mansoni adults, eggs and schistosomula have unique peptidase specificity profiles. *Biochimie* **122**, 99-109, doi:10.1016/j.biochi.2015.09.025 (2016).

45 Käll, L., Krogh, A. & Sonnhammer, E. L. L. Advantages of combined transmembrane topology and signal peptide prediction—the Phobius web server. *Nucleic Acids Research* **35**, W429-W432, doi:10.1093/nar/gkm256 (2007).

46 Reynolds, S. M., Käll, L., Riffle, M. E., Bilmes, J. A. & Noble, W. S. Transmembrane Topology and Signal Peptide Prediction Using Dynamic Bayesian Networks. *PLOS Computational Biology* **4**, e1000213, doi:10.1371/journal.pcbi.1000213 (2008).

47 Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. L. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes11Edited by F. Cohen. *Journal of Molecular Biology* **305**, 567-580, doi:10.1006/jmbi.2000.4315 (2001).

48 Petersen, T. N., Brunak, S., von Heijne, G. & Nielsen, H. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat Meth* **8**, 785-786, doi:10.1038/nmeth.1701 (2011).

49 Hiller, K., Grote, A., Scheer, M., Münch, R. & Jahn, D. PrediSi: prediction of signal peptides and their cleavage positions. *Nucleic Acids Research* **32**, W375-W379, doi:10.1093/nar/gkh378 (2004).

50 Mohri, M., Rostamizadeh, A. & Talwalkar, A. *Foundations of Machine Learning*. (The MIT Press, 2012).

51 Pérez-Sánchez, R., Ramajo-Hernández, A., Ramajo-Martín, V. & Oleaga, A. Proteomic analysis of the tegument and excretory-secretory products of adult Schistosoma bovis worms. *PROTEOMICS* **6**, S226-S236, doi:10.1002/pmic.200500420 (2006).

52 Braschi, S., Borges, W. C. & Wilson, R. A. Proteomic analysis of the shistosome tegument and its surface membranes. *Memórias do Instituto Oswaldo Cruz* **101**, 205-212 (2006).

53 Wang, G. & Dunbrack, J. R. L. PISCES: a protein sequence culling server. *Bioinformatics* **19**, 1589-1591, doi:10.1093/bioinformatics/btg224 (2003).

54 Juszczak, P., Tax, D. & Duin, R. P. in *Proc. ASCI.* 95-102 (Citeseer).

55 Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825-2830 (2011).

56 Japkowicz, N. & Shah, M. *Evaluating Learning Algorithms: A Classification Perspective*. (Cambridge University Press, 2011).

57 Vihinen, M. How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis. *BMC Genomics* **13**, S2, doi:10.1186/1471-2164-13-s4-s2 (2012).

58 Hotez, P. J. *et al.* The Global Burden of Disease Study 2010: Interpretation and Implications for the Neglected Tropical Diseases. *PLOS Neglected Tropical Diseases* **8**, e2865, doi:10.1371/journal.pntd.0002865 (2014).

59 Tebeje, B. M., Harvie, M., You, H., Loukas, A. & McManus, D. P. Schistosomiasis vaccines: where do we stand? *Parasites & Vectors* **9**, 528, doi:10.1186/s13071-016-1799-4 (2016).

60 McManus, D. P. & Loukas, A. Current Status of Vaccines for Schistosomiasis. *Clinical microbiology reviews* **21**, 225-242, doi:10.1128/CMR.00046-07 (2008).

61 Driguez, P. *et al.* Antibody Signatures Reflect Different Disease Pathologies in Patients With Schistosomiasis Due to Schistosoma japonicum. *The Journal of Infectious Diseases* **213**, 122-130, doi:10.1093/infdis/jiv356 (2016).

62 Farnell, E. J. *et al.* Known Allergen Structures Predict Schistosoma mansoni IgE-Binding Antigens in Human Infection. *Frontiers in Immunology* **6**, doi:10.3389/fimmu.2015.00026 (2015).

63 Sammut, C. & Webb, G. I. in *Encyclopedia of Machine Learning* (eds Claude Sammut & Geoffrey I. Webb) 600-601 (Springer US, 2010).

64 Solà, E. *et al.* Evolutionary Analysis of Mitogenomes from Parasitic and Free-Living Flatworms. *PLOS ONE* **10**, e0120081, doi:10.1371/journal.pone.0120081 (2015).

65 Zerlotini, A. *et al.* SchistoDB: a Schistosoma mansoni genome resource. *Nucleic Acids Research* **37**, D579-D582, doi:10.1093/nar/gkn681 (2009).

66 Howe, K. L., Bolt, B. J., Shafie, M., Kersey, P. & Berriman, M. WormBase ParaSite − a comprehensive resource for helminth genomics. *Molecular and Biochemical Parasitology* **215**, 2-10, doi:10.1016/j.molbiopara.2016.11.005 (2017).

67 Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nat Genet* **25**, 25-29, doi:10.1038/75556 (2000).

68 Consortium, G. O. Gene Ontology Consortium: going forward. *Nucleic Acids Research* **43**, D1049-D1056, doi:10.1093/nar/gku1179 (2015).

69 Carbon, S. *et al.* AmiGO: online access to ontology and annotation data. *Bioinformatics* **25**, 288-289, doi:10.1093/bioinformatics/btn615 (2009).

70 Consortium, T. U. UniProt: the universal protein knowledgebase. *Nucleic Acids Research* **45**, D158-D169, doi:10.1093/nar/gkw1099 (2017).

71 Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**, 3389-3402, doi:10.1093/nar/25.17.3389 (1997).

72 Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421, doi:10.1186/1471-2105-10-421 (2009).

73 Ward, N. & Moreno-Hagelsieb, G. Quickly Finding Orthologs as Reciprocal Best Hits with BLAT, LAST, and UBLAST: How Much Do We Miss? *PLOS ONE* **9**, e101850, doi:10.1371/journal.pone.0101850 (2014).

74 Hasan, S. *et al.* SchistoProt: A Highly-Accurate Web Server for Identifying Schistosome-Specific Surface Proteins and Secretory Peptides. *PLOS Neglected Tropical Diseases* (Manuscript submitted for publication).

75 Hasan, S. *et al.* A machine learning approach to classify Schistosoma protein immunoreactivity. (Manuscript in preparation).

76 Mi, H., Muruganujan, A., Casagrande, J. T. & Thomas, P. D. Large-scale gene function analysis with the PANTHER classification system. *Nat. Protocols* **8**, 1551-1566, doi:10.1038/nprot.2013.092 (2013).

77 Mi, H. *et al.* PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Research* **45**, D183-D189, doi:10.1093/nar/gkw1138 (2017).

78 Szklarczyk, D. *et al.* The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Research* **45**, D362-D368, doi:10.1093/nar/gkw937 (2017).

79 Kuhn, M., von Mering, C., Campillos, M., Jensen, L. J. & Bork, P. STITCH: interaction networks of chemicals and proteins. *Nucleic Acids Research* **36**, D684-D688, doi:10.1093/nar/gkm795 (2008).

80 Harn, D. A. *et al.* A protective monoclonal antibody specifically recognizes and alters the catalytic activity of schistosome triose-phosphate isomerase. *The Journal of Immunology* **148**, 562-567 (1992).

81 Figueiredo, B. C. *et al.* Schistosome Syntenin Partially Protects Vaccinated Mice against Schistosoma mansoni Infection. *PLOS Neglected Tropical Diseases* **8**, e3107, doi:10.1371/journal.pntd.0003107 (2014).

82 Aye, I. L. M. H., Singh, A. T. & Keelan, J. A. Transport of lipids by ABC proteins: Interactions and implications for cellular toxicity, viability and function. *Chemico-Biological Interactions* **180**, 327-339, doi:10.1016/j.cbi.2009.04.012 (2009).

83 Horn, M. *et al.* Trypsin- and Chymotrypsin-Like Serine Proteases in Schistosoma mansoni – 'The Undiscovered Country'. *PLOS Neglected Tropical Diseases* **8**, e2766, doi:10.1371/journal.pntd.0002766 (2014).

84 Quezada, L. A. L. & McKerrow, J. H. Schistosome serine protease inhibitors: parasite defense or homeostasis? *Anais da Academia Brasileira de Ciências* **83**, 663-672 (2011).

85 Patocka, N., Sharma, N., Rashid, M. & Ribeiro, P. Serotonin Signaling in Schistosoma mansoni: A Serotonin–Activated G Protein-Coupled Receptor Controls Parasite Movement. *PLOS Pathogens* **10**, e1003878, doi:10.1371/journal.ppat.1003878 (2014).

86 Girardini, J. E., Khayath, N., Amirante, A., Dissous, C. & Serra, E. Schistosoma mansoni: Ferredoxin-NADP(H) oxidoreductase and the metabolism of reactive oxygen species. *Experimental Parasitology* **110**, 157-161, doi:10.1016/j.exppara.2005.02.011 (2005).