MINING FOR CONSERVED MOTIFS AND SIGNIFICANT FUNCTIONS IN *S. MANSONI* CERCARIAL SECRETIONS

Amy L. Schmidbauer

Submitted to the faculty of the School of Informatics
in partial fulfillment of the requirements
for the degree
Master of Science
in Bioinformatics,
Indiana University

December 30, 2006

Accepted by the Faculty of Indiana University, in partial
fulfillment of the requirements for the degree of Master of Science
in Bioinformatics.

(Committee Chair's signature)________________________
Sean D. Mooney, Ph.D., Chair

Master's Thesis
Committee

(Second member's signature)_________________________
Xiaoman Shawn Li, Ph.D.

(Third member's signature)__________ _______________
William J. Sullivan, Ph.D.

© 2006
Amy L. Schmidbauer
All Rights Reserved

ACKNOWLEDGMENTS

This project would not have been possible without the guidance and support of many people including faculty, advisors, colleagues, friends and family. I am extremely grateful to my advisor, Dr. Sean Mooney, for welcoming me into his laboratory as a graduate student, and for providing, not only computing resources, but continued support, suggestions, and guidance as, what started out as an independent study project, grew into what became this thesis. I extend my sincere appreciation to Dr. Giselle Knudsen, an honorary member of my thesis committee, for the original project inspiration, for her enthusiastic encouragement, insight, and direction throughout the project, and for her thoughtful review of this thesis. I would like also like to extend a heartfelt thank you to Dr. William Sullivan and Dr. Xiaoman Li for their willingness to lend their time to reviewing my thesis and for the insightful feedback they provided. For statistical expertise and support I would like to extend my deepest appreciation to Dr. Eunseog Eun. Thank you also to Dr. Snehasis Mukhopadhyay for advice provided in the early stages of this work. For assistance with various software installations, I would like to thank Brandon Peters and Dustin Reynolds, and for assistance early in the project with BioPerl, I would like to thank Jessica Dantzer.

It is my pleasure to thank my employer, Dow AgroSciences LLC, Indianapolis, IN, for providing computing resources and funding the majority of my graduate work. My sincere gratitude is extended to Dr. Neil Kirby of Dow AgroSciences for his flexibility and support throughout my graduate studies. Also, I would like to thank my colleagues Dr. Ignacio Larrinua, for consulting on various issues, and Mindi Dippold, for her assistance with Perl.

Finally, I am very grateful to my friends and family, and especially Terry, for their support and encouragement throughout this journey.

ABSTRACT

Amy L. Schmidbauer

MINING FOR CONSERVED MOTIFS AND SIGNIFICANT FUNCTIONS IN *S. MANSONI* CERCARIAL SECRETIONS

Schistosomiasis is a disease caused by a parasitic flatworm of the genus *Schistosoma.* It infects an estimated 200 million people and 165 million head of livestock worldwide. There is medical interest in characterizing the parasitic proteins that interact with the human host for either the development of vaccines or the identification of drug targets. The cercarial secretome and adult tegument sub-proteome of *S. mansoni* have both been recently published (Knudsen, Medzihradszky *et al.* 2005) (van Balkom, van Gestel *et al.* 2005). As secretome proteins are secreted extracellularly, and tegument sub-proteome proteins are anchored in the cellular membrane, we hypothesize that both sets of proteins employ similar secretion machinery and mechanisms. Motivated by the discovery in the malarian parasite, *Plasmodium falciparum*, of conserved sequence motifs that are required for export downstream of N-terminal signal sequences (Hiller, Bhattacharjee *et al.* 2004), *S. mansoni* secretory and tegumental proteins were analyzed for conserved motifs using recursive iterations of MEME and MAST. To compliment the conserved motif analysis, an automated workflow to process InterProScan functional domain and GO annotation data, that employs statistical methods for determining significant functions, was developed. A conserved motif, enriched in the mechanically-induced vesicle secretion proteins, was elucidated and insight was gained into both the functions of proteins found to contain the motif, as well as the effects of different cercarial secretion induction methods. A hypothesis of the secretion model empoloyed by the invading parasite was generated.

TABLE OF CONTENTS

I. INTRODUCTION 1
  A. Introduction of Subject 1
  B. Importance of Subject 1
  C. Knowledge Gap 2
II. BACKGROUND 6
  A. Related Research 6
  B. Current Understanding 12
  C. Research Question 26
  D. Intended Research Project 26
III. METHODS 27
  A. Materials: Software and Hardware 27
  B. Samples and Subjects 28
  C. Procedures and Interventions 30
  D. Statistical Analysis 32
  E. Expected Results 32
  F. Alternate Plans 33
IV. RESULTS 36
  A. Introduction 36
  B. Important Highlights 36
  C. Specific Findings 37
  D. Summary 112
V. CONCLUSION 112
  A. Overview of Significant Findings 112
  B. Consideration of Findings in Context of Current Knowledge 114
  C. Theoretical Implications of the Findings 115
VI. DISCUSSION 116
  A. Limitations of the Study 116
  B. Recommendations for Further Research 116
APPENDICES 118
  Appendix A: Summary of Sample Data Sets 118
  Appendix B: Control Data Set 126
  Appendix C: Sequence Homology Within Sample Sets 132
  Appendix D: Sequence Homology Between Sample Sets 136
  Appendix E: ProP Propeptide prediction results 140
  Appendix F: Gene Ontology Annotation 152
  Appendix G: Protein Domain Annotation 160
  Appendix H: GO Entropy Calculations 171
  Appendix I: Protein Domain Entropy Calculations 177
  Appendix J: Perl And Matlab Scripts 182
  Appendix K: Permission To Use Figures 193
REFERENCES 198
CURRICULUM VITAE

LIST OF FIGURES

Figure 1. Praziquantel (left) and Oxamniquine (right) 4
Figure 2. Life cycle of schistosomes....................14
Figure 3. Scanning electron micrograph of a *S. mansoni* cerariae....................15
Figure 4. Cross section of a *S. mansoni* cercaria 16
Figure 5. Anterior organ of a *S. mansoni* cercaria....................17
Figure 6. Components of the *S. mansoni* tegument....................18
Figure 7. Workflow #1 34
Figure 8. Workflow #2 35
Figure 9. Comparison of 3 Sample Sets....................38
Figure 10. *S. mansoni* amino acid frequencies....................49
Figure 11. MEME #1....................53
Figure 12. MAST #1 high scoring sequences 55
Figure 13. MAST #1 motif diagram....................56
Figure 14. MEME #2....................58
Figure 15. MAST #2 high scoring sequences 60
Figure 16. MAST #2 motif diagram....................61
Figure 17. MAST #3 high scoring sequences 63
Figure 18. MAST #3 motif diagram....................64
Figure 19. Multiple sequence alignment of sample proteins with the predicted xKxGE motif 67
Figure 20. Distribution of proteins found to contain xKxGE motif.....................69
Figure 21. Top five BLAST hits for myosin light chain AAD41591 70
Figure 22. WebLogo of motif found in *S. mansoni* sample proteins....................71
Figure 23. WebLogo of motif in *P. falciparum.*....................72
Figure 24. Venn diagram of similarity across sample proteins 80
Figure 25. Protein domains corresponding to vesicle-specific proteins....................83
Figure 26. GO terms corresponding to vesicle-specific proteins.....................84
Figure 27. Protein domains corresponding to secretion-specific proteins 85
Figure 28. GO terms corresponding to secretion-specific proteins....................86
Figure 29. Protein domains and GO terms corresponding to tegument-specific proteins......87
Figure 30. Phylogenetic tree of vesicle proteins....................88
Figure 31. Phylogenetic tree of secretion proteins....................89
Figure 32. Phylogenetic tree of tegument proteins....................90
Figure 33. InterProScan raw data file example....................101
Figure 34. % of sequences corresponding to GO annotations from category biological process....................103
Figure 35. % of sequences corresponding to GO annotations from category cellular component 104
Figure 36. % of sequences corresponding to GO annotations from category molecular function (I)....................105
Figure 37. % of sequences corresponding to GO annotations from category molecular function (II) 106
Figure 38. % of sequences corresponding to InterPro protein domains....................109

LIST OF TABLES

Table 1. Signal peptide predictions..39
Table 2. MEME #1 input sequences..52
Table 3. MEME #2 input sequences..57
Table 4. Sample proteins containing xKxGE motif...65
Table 5. Proteins containing predicted conserved motif vs. protein domains 76
Table 6. Sample sequences containing predicted conserved motif vs. GO terms........77
Table 7. Summary of motif-containing vesicle and secretion proteins. 78
Table 8. GO terms and protein domains associated with phylogenetic groups...........91
Table 9. GO terms with highest entropy..108
Table 10. Protein domains with highest entropy values 111

GLOSSARY

1D SDS PAGE- One dimensional sodium dodecyl sulphate polyacrylamide gel electrophoresis. A technique for separating large numbers of proteins by loading them onto a gel and applying an electric current.

2D gel electrophoresis- Two dimensional polyacrylamide gel electrophoresis. A technique for separating large numbers of proteins by loading them onto a gel, applying an electric current to separate by isoelectric point, then, in the perpendicular direction, separating them by molecular weight.

acetabular glands- *S. mansoni* possesses 2 sets of acetabular glands, one set posterior and one set anterior to the acetabulum (ventral sucker), which produce mucus and enzymes that are important in skin penetration.

acetabulum- Ventral sucker of *S. mansoni.* Plays an active role in skin exploration and penetration.

BAC- Bacterial artificial chromosome. A DNA construct, based on a fertility plasmid, used for transforming and cloning in bacteria, usually *E. coli*. Its usual insert size is 150 kbp, with a range from 100 to 300 kbp.

Bayes optimal classifier- Statistical classifier for predicting class membership probabilities, based on Bayesian theorem which uses prior and posterior probabilities.

bioinformatics- The science of informatics as applied to biological research. Informatics is the management and analysis of data using advanced computing techniques.

BLAST- Basic Local Alignment Search Tool. An algorithm for performing similarity searches through local alignments.

blood fluke- A term used to refer to *Schistosoma*, a trematode flatworm.

cDNA- Complimentary DNA. Single-stranded DNA that has been synthesized from an mRNA template by reverse transcriptase.

centromere- The cinched "waist" of the chromosome essential for the division and the retention of the chromosome in the cell. Uniquely specialized region of the chromosome to which spindle fibers attach during cell division.

cercariae (plural, singular = cercaria)- the invasive larval stage of the blood fluke.

CLUSTALW- A general purpose program for performing multiple sequence alignments of DNA or protein sequences.

contig- Short for contiguous. Refers to a contiguous set of overlapping DNA sequences in the context of a sequencing project. Also referred to as an EST assembly, tentative consensus sequence, or gene cluster.

definitive host- The host in which the sexual reproduction of a parasite takes place

erythrocytes- Blood cells.

escape glands- schistosome cercarial glands that excrete their contents when the cercaria emerges from its snail host.

EST- Expressed Sequence Tag. Short, single-pass reads from mRNA that are usually produced in large numbers in the context of sequencing projects. Represent portion of DNA that is expressed.

EST assemblies- see “contig”.

ESTScan- Program for detecting ORFs (open reading frames); able to find and correct sequence errors resulting in frame shifts within the coding sequence.

e-value- Expectation value. For a given score, (e.g. BLAST) the number of hits in a database search that are expected to be seen by chance. Takes into account the size of the database. The lower the E-value, the more significant the score.

expectation maximization- A technique for estimating hidden parameters. Used by MEME.

fibrosis- Formation of scar-like tissue.

FISH- Fluorescent In Situ Hybridization. A laboratory technique used to determine how many copies of a specific segment of DNA are present or absent in a cell in which a special region of a chromosome is stained with a dye that emits colored light when exposed to ultraviolet light.

FrameFinder- Program for detecting ORFs.

gene clusters- See “contig”.

Gibbs sampling strategy- Local multiple sequence alignment method that uses a stochastic algorithm.

glycocalyx- A sticky layer of a cell wall that consists of proteins and/or polysaccharides and functions to maintain osmotic pressure in the organism.

Golgi apparatus- Membrane-bound organelle in eukaryotic cells where the proteins and lipids made in the endoplasmic reticulum are modified and sorted.

granuloma- The formation of a nodule as a result of inflammation.

head gland- One of the glands of *Schistosoma mansoni*. Postulated to contain secretory bodies of proteins that function in the adjustment of the schistosomulae post-penetration.

HMM- Hidden Markov Model. A statistical model where the system being modeled is assumed to be a Markov process (a simple stochastic process in which the distribution of future states depends only on the present state and not on how it arrived in the present state) with unknown parameters, and the challenge is to determine the hidden parameters, from the observable parameters, based on this assumption. The extracted model parameters can then be used to perform further analysis, for example for pattern recognition applications.

homology- Pertaining to phylogeny. Particular features (e.g. genes or proteins) in different individuals that are descended from the same feature in a common ancestor.

immunogenic- Possessing the ability to elicit an immune response.

immunogenicity- The ability of an antigen or vaccine to stimulate an immune response.

information content- Referring to the amount of information a particular motif found by MEME contains. Highly conserved positions in the motif have high information; positions where all letters are equally likely have low information.

intermediate host- Host that serves as a temporary but essential environment for the completion of a parasite's life cycle.

InterPro- A federated database of profiles, patterns, and HMMs for detecting protein domains.

LC-MS/MS- Liquid Chromatography/Mass Spectrometry/Mass Spectrometry. A technique that combines the solute separation power of HPLC, with the detection power of a mass spectrometer for identification of proteins.

lumen- The cavity or channel within a tube or tubular organ such as a blood vessel or the intestine.

malaria- An infective disease caused by a protozoan parasites that is transmitted through the bite of an infected *Anopheles* mosquito.

MAST- Motif Alignment Search Tool. A tool for searching sequence databases for motifs found by MEME.

Matlab- A commercial package (The MathWorks) which operates as an interactive programming environment, especially good for mathematical operations and working with matrices.

MEME- Multiple Expectation Maximization for Motif Elicitation. A tool for finding sequence motifs.

metazoan- Multicellular organism.

miracidia (plural, singular = miracidium)- the small free-swimming larvae of *S. mansoni* that hatch from eggs present in host feces, swim to find their freshwater snail intermediate host, and penetrate snails tissues where they develop into sporocyts.

MS/MS- Mass Spectrometry/Mass Spectrometry. Tandem mass spectrometry. A method by which peptide sequences are solved from the mass-to-charge (m/z) fragmentation patterns of ions produced from collision-induced-dissociation.

multiple sequence alignment- A bioinformatics tool that compares multiple DNA or amino acid sequences and aligns them to highlight their similarities.

oral sucker- The part of the *S. mansoni* cercaria that works its way into its host's skin.

ORESTES- Open Reading frame Expressed Sequence Tags. A strategy for preferentially generating ESTs of the central, and thus most informative portion of the transcript. Also frequently identifies less abundant mRNAs.

parasitophorous vacuolar membrane- the membrane surrounding the vacuole that is created by plasmodia for survival within host erythrocyte cells.

parasitophorous vacuole- The vacuole that is created by plasmodia for survival within host erythrocyte cells.

pathogenicity- Capacity to cause disease.

PEXEL motif- Plasmodium Export Element. A five amino acid conserved sequence motif found in the malaria parasite, *P. falciparum,* that allows it to survive and multiply by exporting multiple remodeling and virulence proteins across the membrane of the parasitophorous vacuole into human erythrocytes.

PfEMP1- Plasmodium falciparum erythrocyte membrane protein 1. An important virulence protein in the malaria parasite.

PSPM- Position Specific Probability Matrix. Used by MEME to represent the observed frequency of each letter in any motif it finds. Columns correspond to letters in the alphabet, rows correspond to position in the motif.

PSSM- Position Specific Scoring Matrix. A log-odds matrix calculated by taking 100 times the log (base 2) of the ratio *p/f* at each position in the motif, where *p* is the probability of a particular letter at that position in the motif, and *f* is the background frequency of the letter. Used by MAST to detect occurrences of the motif in sequence databases. Columns correspond to letters in the alphabet, rows correspond to position in the motif.

postacetabular glands- Part of *S. mansoni's* cercarial morphology. Three pairs of glands are located posterior to the acetabulum (ventral sucker). Vesicle secretions produced by these glands contain various enzymes, including proteases and high levels of calcium, making them important in skin penetration.

preacetabular glands- Part of *S. mansoni's* cercarial morphology. Two pairs of glands located anterior to the acetabulum (ventral sucker). Produce mucus and help cercariae adhere to surfaces.

PRINS- Primed In Situ Hybridization. An alternative method to traditional *in situ* hybridization for the identification of chromosomes in metaphase spreads or interphase nuclei. Denatured DNA is hybridized to short DNA fragments, or designed oligonucleotides and then primer extension is carried out with the appropriate polymerase in the presence of labeled nucleotides.

proteome- The total complement of proteins expressed by an organism.

proteomics- The study of the proteome, employing methods that separate, quantify and identify proteins.

protozoan- Single-celled eukaryote.

PFGE- Pulsed Field Gel Electrophoresis. A gel electrophoretic method for the separation of megabase fragments of DNA based on continuous alteration of the angle at which the electrical field is applied.

p-value- The p-value indicates the probability that the result obtained in a statistical test is due to chance rather than a true relationship between measures. Small p-values indicate that it is very unlikely that the results were due to chance.

reporter gene- A gene whose gene product is easily detected (e.g. through the use of fluorescent labeling).

RT-PCR- Reverse Transcriptase Polymerase Chain Reaction. Method for amplifying DNA or RNA segments using heat-resistant enzymes. First, complementary DNA (cDNA) is made from an RNA template, using a reverse transcriptase enzyme, and then some of it is used in a PCR reaction to produce large quantities.

schistosome: The blood fluke or flatworm.

schistosomiasis- The disease caused by species of *Schistosoma*.

schistosomula (plural, singular = schistosomules)- One of the life cycle stages of *Schistosoma*. Cercaria are transformed into schistosomula after they penetrate the skin.

secretome- The set of soluble proteins secreted into the extracellular environment of a cell.

secretory vesicles- Membrane bound structure derived from the Golgi apparatus, containing material to be released from the cell.

signal peptide- A short peptide sequence (15-60 amino acids long) that directs the post translational transport of a protein. Some signal peptides are cleaved from the protein by signal peptidase after the proteins are transported. Signal peptides may also be called targeting signals or signal sequences.

signal sequence trap method- A method for selecting cDNAs encoding secreted and surface proteins with N-terminal signal peptides. cDNAs are fused with a signal sequence-deficient reporter gene with the expectation that those cDNAs containing a signal sequence will result in surface expression of the protein produced.

SignalP- A program for predicting the presence of signal peptides and location of signal peptide cleavage sites in amino acid sequences from different organisms, based on pattern recognition of known motifs for signal peptidase I found in higher eukaryotes.

singletons- ESTs that are not contained in an assembly.

sporocysts- One of the life cycle stages of *Schistosoma.* Miracidia transform into sporocysts after they penetrate the tissues of their intermediate host (the snail). Sporocysts then multiply asexually to produce hundreds of thousands of cercariae.

tegument- The outer membrane of the blood fluke. The tegument in *Schistosoma* is uniquely 3-layered.

tentative consensus sequences- See “contig”.

Th1-type response- Two types of T-helper cells boost the immune attack. Th1-type cells stimulate macrophages and natural killer cells, which directly attack microbes that replicate in the body’s cells. This type of response is called cellular immunity.

Th2-type response- Two types of T-helper cells boost the immune attack. Th2 cells attack foreign matter too large to be killed by macrophages or natural killer cells, by preferentially stimulating B cells to produce antibodies. This type of response is called humoral immunity.

UNDP- United Nations Development Program.

unsupervised learning- A method of machine learning where a model is fit to observations. Does not involve any prior knowledge, unlike supervised learning. A data set of input objects is gathered and treated as a set of random variables, for which a joint density model is then built.

VTS- Vacuolar Transport Sequence. An eleven amino acid conserved sequence motif found in the malaria parasite, *P. falciparum,* that allows it to survive and multiply by exporting multiple remodeling and virulence proteins across the membrane of the parasitophorous vacuole into human erythrocytes. (The PEXEL motif is contained within the VTS.)

ventral sucker- See “acetabulum”.

weight matrix- A flexible type of quantitative motif descriptor that contains weights or scores for each amino acid in every position.

Weka- Data mining software written in Java. Collection of machine learning algorithms for data mining tasks.

WHO TDR- World Health Organization Special Program for Research and Training in Tropical Diseases.

whole genome shotgun sequencing- A method of sequence where all the DNA is first broken into fragments and the fragments are then sequenced at random and assembled together by looking for overlaps.

I. INTRODUCTION

A. Introduction of Subject

Schistosoma mansoni is a parasitic flatworm that causes the disease schistosomiasis and infects an estimated 200 million people worldwide. Schistosomiasis is generally considered to rank second behind malaria among parasitic/infectious diseases in global importance. It is endemic to 76 countries, mostly underdeveloped, and is usually associated with water resource development projects that expand the habitat of the parasite's intermediate host, the freshwater snail (Johnston, Blaxter *et al.* 1999). There is medical interest in characterizing the parasitic proteins that interact with the human host for either the development of vaccines or identification of drug targets. The proteins that the larval stage of the parasite secretes extracellularly just prior to infection of the definitive human host have been recently isolated and identified (Knudsen, Medzihradszky *et al.* 2005). Proteins contained in the adult tegument, the metabolically active outer surface of the parasite, have also recently been isolated and identified (van Balkom, van Gestel *et al.* 2005). The recent discovery of highly conserved sequence motifs in proteins exported by the protozoan parasite *Plasmodium falciparum*, which causes malaria, were the motivation of this research (Hiller, Bhattacharjee *et al.* 2004). We hypothesize that a conserved motif in proteins that are either secreted extracellularly or anchored into the tegument underlies a common export mechanism. The goals of this project are to elucidate a conserved sequence motif across *S. mansoni* secreted proteins and to functionally annotate the proteins to hypothesize about the underlying mechanism of secretion used by the invading parasite.

B. Importance of Subject

According to the World Health Organization's Tropical Disease Research Unit (WHO TDR), approximately 600 million people worldwide are at risk for schistosomiasis infection, 200 million people are currently infected, 20 million severely. The majority of infections occur in under-developed countries (TDR 2002). Reported mortality estimates range from 11,000 per year (TDR 2002) to 500,000 per year (Johnston, Blaxter *et al.* 1999). It has been given a Category 2 rating by the World Health Organization, meaning that a control strategy exists but has not resulted in a persistent reduction in the disease burden (TDR 2002). Reinfection is common and can lead to liver, intestine, and urinary tract damage and can stunt growth and cognitive development (Remme, Blas et al. 2002). The parasite also infects approximately 165 million head of livestock (Chitsulo, Engels et al. 2000).

C. Knowledge Gap

Two drugs introduced in the 1970s are still the only ones commercially available for the treatment of schistosomiasis: praziquantel and oxaminquine (Cioli and Pica-Mattoccia 2005). Oxamniquine is only effective against *S. mansoni*, limiting its use to South America where *S. mansoni* is the only species present. Praziquantel is a broader spectrum drug that has been used in other parts of the world. Market competition has driven down the price of praziquantel, but not oxamniquine, due to its geographical limitations. This has resulted in countries even in South America decreasing their use of oxamniquine in favor of praziquantel. This has effectively made praziquantel the only drug currently available against the disease (Cioli and Pica-Mattoccia 2005). Praziquantel is effective in a single dose, but given that infections occur mainly in under-developed countries (85% in sub-Saharan Africa) and that reinfection is common, the cost of $0.40 per treatment is prohibitive (Cioli and Pica-Mattoccia 2005). Massive funding from the Gates Foundation is allowing implementation of a program called the “Schistosomiasis Control Initiative”, which will provide praziquantel treatments to a few select African nations, frequently to school-aged children (Cioli and Pica-Mattoccia 2005). But full-blown resistance to praziquantel could occur at any time (Chitsulo 2005), so a schistosomiasis vaccine, as well as new lead chemotherapeutics, need to be pursued.

Both drugs are chemotherapeutics whose modes of action are not completely understood. Both drugs result in selective toxicity to the adult worm. Their structures are shown in Figure 1 below. It is known that oxamniquine’s mode of action is dependent upon the presence of the hydroxymethyl group (Cioli and Pica-Mattoccia 2005). Some interesting features of oxamniquine activity include: 1) schistosomes have a delayed reaction to it (5-7 days after a 30 minute in vitro exposure), 2) the length of exposure can be as short as 15 minutes for the drug to be effective, 3) it is more effective against male worms than female worms, and 4) it has little activity against immature schistosomes (Cioli and Pica-Mattoccia 2005). Resistance to oxamniquine has been proven both in the laboratory (Rogers and Bueding 1971) and in the field (Katz, Dias et al. 1973) and it has shown to be complete (Pica-Mattoccia, Dias et al. 1993). This resistance does not, however, spread throughout the population, as cases of it have been confined and random (Cioli and Pica-Mattoccia 2005). It has been proven that resistance to the drug is a recessive trait (Cioli, Pica-Mattoccia et al. 1992), which has led to the suggestion that resistant schistosome are missing some function that activates the drug (Cioli and Pica-Mattoccia 2005). This missing function has been proven to be the enzyme sulfotransferase (Cioli and Pica-Mattoccia 2005). It has been shown that a particular sulfotransferase is specific to *S. mansoni,* and is necessary to confer sensitivity to the drug oxamniquine. Sulfotransferases function by transferring a sulfate group from a donor molecule (phosphoadenosine-phosphosulfate, PAPS) to the hydroxyl group of an acceptor substrate, oxamniquine, in this case. This creates an unstable sulfate ester which spontaneously dissociates, leading to an electrophile (positively charged molecule) capable of alkylating (forming covalent bonds with) other molecules, such as, in this case, schistosome DNA and other macromolecules. The sulfotransferase gene has not yet been cloned in *S. mansoni*. Schistosomes without a functional sulfotransferase have a lower fitness, as indicated by various measure of the vitality of their life cycle. This decreased fitness, in addition to the fact that the gene is recessive, is proposed as an explanation for why oxamniquine resistance does not spread throughout a population (Cioli and Pica-Mattoccia 2005).

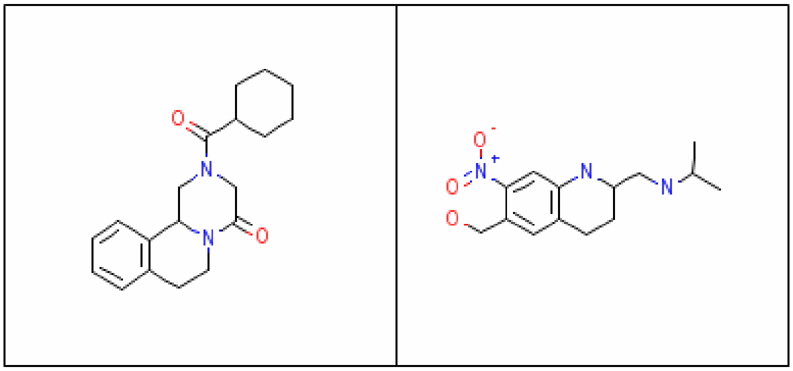


Figure 1. Praziquantel (left) and Oxamniquine (right)

Praziquantel is effective not just against all species of schistosomes, but also against other trematodes and human and veterinary cestodes. The effectiveness of the drug is significantly reduced in animals that are immune-compromised (Sabah, Fletcher et al. 1985), though this has not yet been evidenced in humans (Karanja, Boyer et al. 1998). The drug has little effect on immature schistosomes (Gonnert and Andrews 1977), and the height of effectiveness against mature schistosomes is reached at two points: very early, -1 to +1 days from infection, and relatively late, 7 weeks post infection. The biggest effect of the drug is a huge increase in the amount of intracellular calcium in the parasite (Pax, Bettett et al. 1978) if calcium is present in the medium. The cause of this influx is not known, but it has been shown not be caused be alteration of either ATPases that pump calcium out of cells (Nechay, Hillman et al. 1980) or voltage-gated calcium channels (Fetterer, Pax et al. 1980). This huge increase in calcium causes severe paralysis of the parasite's musculature, causing the schistosome suckers to release their hold on the inner wall of blood vessels and the schistosome to migrate to the liver. Another significant effect caused by praziquantel is massive alterations of the parasite's tegument, also dependent on the presence of calcium. This increases the number of schistosome antigens present at the parasite surface, some of which have been identified and characterized (Doenhoff, Sabah et al. 1987) (Sauma and Strand 1990), which exposes antigens to the host immune response.

A pyrazio-isoquinoline ring system containing an asymmetric carbon atom that leads to two enantiomers referred to as *dextro*- and *laevo*- praziquantel form the base of praziquantel structure. Despite the fact that the *dextro-* form is almost inactive while the *laevo-* form possesses most of the anti-schistosomal activity, most commercial preparations are a 50:50 mixture because isolating the *laevo-* form is an expensive process. Hypotheses regarding the mode of action include that praziquantel interacts with membrane phospholipids to change the permeability of the lipid bilayer (Harder, Goossens et al. 1988) (Schepers, Brasseur et al. 1988), and that it interacts somehow with glutathione S-transferase (GST) based on the presence of a pocket in the three-dimensional structure of GST that could fit praziquantel (McTigue, Williams et al. 1995). These hypotheses have been disproven, however. More recently it has been hypothesized that beta subunits of calcium-binding proteins are involved in praziquantel activity, based on them containing subsequences that are unique to schistosomes (Kohn, Anderson et al. 2001).

There is currently no vaccine available for schistosomiasis. Challenges involved in studying schistosomes include: 1) the lack of an animal model that adequately reflects the human response, 2) maintaining the life cycle of *Schistosoma* in the laboratory (*S. mansoni* is used most often because it is the easiest to maintain in the laboratory), 3) finding a vaccine with broad spectrum activity against all five species of *Schistosoma* that infect human (or at least the three main species), and 4) the fact that the parasite goes through three life cycle changes within its human host means that the antigens presented to the host are always changing. There is reason to believe that a vaccine can be developed, however, as there is ample evidence that humans acquire some level of resistance after recurrent schistosome infection (Bergquist 1998). In the mid-90s the WHO selected six promising anti-larval or anti-worm vaccine candidate antigens for testing in mouse and rat including: glutathione-S-transferase (Sm28) and the muscle protein paramyosin (Sm97), both found in the schistosomula and adult stages; triose phosphate isomerase (TPI), a membrane antigen (Sm23), and myosin heavy chain (rIrV5), all found in all stages; and the antigenic membrane protein fatty acid-binding protein (Sm14), found only in the schistosomula stage (Mountford and Jenkins 2005). The target of consistently inducing 40% protection was not reached, however, most likely due to instability of formulations (Bergquist 1998). The ability to scale up production according to GMP standards has turned out to be an important criterion in assessing vaccine candidates (Bergquist 2004).

New candidate antigens are needed and "expectations of finding additional candidates through mining the expanding genomics databases and through proteomics analysis are justifiable" (Bergquist 2004). The underlying mechanisms of human resistance to schistosomiasis and the immune responses elicited by the various life cycle stages or particular antigens are not well understood. For example, some antigens have been shown to require a Th2-type response, others a Th1-type response (Bergquist 2004). There is also disagreement in the research community about the desired type of immune response that a vaccine should achieve. This lack of agreement and understanding has lead to different adjuvants being selected for delivery of antigens and the selected adjuvant itself has an effect on the immune responses elicited (Mountford and Jenkins 2005). Control strategies now seem to be aiming towards development of vaccines that provide partial protection that can be complemented with chemotherapeutic drugs (Bergquist 2004).

II. BACKGROUND

A. Related Research

S. mansoni

The Schistosoma Genome Network (SGN) is one of the Parasite Genome Initiatives of the WHO. It is an international collaboration of laboratories that has received support from the WHO/UNDP/World Bank Special Program for Research and Training in Tropical Diseases (TDR) since 1994 (Johnston, Blaxter et al. 1999). The financial support for this project provided by WHO is motivated by the desire to:

> "1) increase knowledge of parasite molecular biology, especially with respect to mechanisms of drug resistance, antigenic variation, and genetic diversity;
> 2) identify genes with key cellular functions that could represent new drug targets and to identify antigens with diagnostic and/or vaccine potential;
> 3) where practical, develop a physical map of the parasite's genome and/or other library resources for distribution to the wider research community; and
> 4) curate, analyze, and disseminate parasite genome data" (Johnston, Blaxter et al. 1999).

The first *S. mansoni* genes were cloned in the early 1980s, but by the early 1990s there was still little progress made towards elucidating the *S. mansoni* genome: only a few more than one hundred cDNAs had been deposited in GenBank (Ohler and Niemann 2001). Initial efforts of the SGN focused on generating large numbers of expressed sequence tags (ESTs). In 1994 partial sequencing of *S. mansoni* cDNA clones generated 607 ESTs, of which 16% had been previously identified, 22% showed homology to other organisms, and 33% showed no significant homology to other organisms (Ohler and Niemann 2001). This was one of the first large-scale studies to illuminate new *S. mansoni* genes.

A large-scale effort across multiple labs in the State of São Paulo was initiated in 2001 to sequence *S. mansoni* EST clones generated by the ORESTES low-stringency RT-PCR amplification technique (Williams and Pierce 2005). The central portion of messages containing the more conserved, function-defining coding regions are amplified in this technique through the use of arbitrary primers and low-stringency RT-PCR (Fietto, DeMarco et al. 2002). 163,000 ESTs from six developmental stages of *S. mansoni* were sequenced, resulting in 31,000 gene clusters, covering an estimated 92% of an estimated 14,000 expressed genes (Verjovski-Almeida, DeMarco et al. 2003). Approximately 124,000 ESTs, 11,000 contigs, and 15,000 singletons resulting from this effort are available to the public through the *S. mansoni* Assembled ESTs database (April 2002, URL:http://cancer.lbi.ic.unicamp.br/schisto6/).

Additional EST sequencing efforts have resulted in significant progress towards identifying both *S. mansoni* and *S. japonicum* genes. *S. mansoni* EST sequences have been assembled, archived and made available to the public at The Institute for Genomic Research's (TIGR) *S. mansoni* Gene Index (SMGI) database (The TIGR Gene Index Databases, The Institute for Genomic Research, Rockville, MD 20850, URL: http://www.tigr.org/tdb/tgi, (Quackenbush, Liang et al. 2000) . It contains approximately 138,000 ESTs, 13,000 tentative consensus sequences, and 21,000 singletons (version 5, January 22, 2005).

In June 2003, the Sanger Institute initiated a 5x whole genome shotgun sequencing effort and clusters were created using a combined set of Sanger and TIGR sequences, giving a total of 8x coverage of the genome (El-Sayed, Bartholomeu et al. 2004). Sanger's *S. mansoni* GeneDB (version 3.0, October 14, 2005) contains a total of 50, 376 contigs. A project to complete sequencing of a 1.1 Mb *S. mansoni* BAC contig has also been initiated at The Sanger Institute.

The SGN's focus has been on *S. mansoni* mainly. This is evident in the comparative numbers of ESTs contained in NCBI's EST database, dbEST. As of June 12, 2006, it database contains 158,841 *S. mansoni,* 97,526 *S. japonicum*, and 6 *S. haematobium* ESTs (URL: http://www.ncbi.nlm.nih.gov/dbEST/index.html). There is, however, growing understanding that the different schistosome species vary in many ways including their range of definitive hosts, egg production, pathogenicity, and immunogenicity (Le, Blair et al. 2002). This has motivated various projects that focus on other species including the sequencing of 43,000 ESTs from the egg and adult stages of *S. japonicum* as part of a cooperative project between the Chinese National Human Genome Center at Shanghai (CHGCS), Shanghai Institute of Parasitology, Chinese Academy of Medical Science, and Shanghai Second Medical University. 13,000 gene clusters resulted, of which 35% had no homology to known genes and 75% had not been previously reported (Hu, Yan et al. 2003). All ESTs, cDNAs, and EST assemblies are available through the Chinese Human Genome Center at Shanghai (URL: http://schistosoma.chgc.sh.cn/sj-old/index.htm) and have also been deposited in GenBank. The availability of this data set is a great asset in the study of *S. mansoni*, as the two organisms are closely related.

Very little work has been done to date on *S. haematobium*. But recently a collaboration has begun at The Sanger Institute to generate 15,000 ESTs from both *S. haematobium* and *Fasciola hepatica. F. hepatica* will be useful in comparative studies because it is also a platyhelminth digean that is phylogenetically close to *Schistosoma* and is an important human and veterinary pathogen. This EST data will be available at Sanger's Sm GeneDB (http://www.genedb.org/).

The signal sequence trap method was first developed and described in the early 1990s as a method for selecting cDNAs encoding secreted and surface proteins with N-terminal signal peptides. cDNAs are fused with a signal sequence-deficient reporter gene with the expectation that those cDNAs containing a signal sequence will result in surface expression or secretion of the protein produced. It is an especially effective technique for studying genomes of organisms not yet sequenced. The first application of this technique for identifying surface and secreted proteins from any pathogenic organism was performed in *S. mansoni*, where a protein kinase, tetraspanins similar to human cell surface antigens, and previously characterized *S. mansoni* surface proteins were identified from the adult stage blood fluke (Smyth, McManus et al. 2003). Further work applying this method to cDNAs expressed during various *S. mansoni* life cycle stages revealed additional cDNAs containing a signal sequence as well as one encoding a seven transmembrane receptor (Pearson, McManus et al. 2005). A bioinformatics analysis of 100 signal peptides each from 4 different species including *Schistosoma*, human, *Escherichia coli*, and *Nippostrongylus brasiliensis* (a parasitic nematode) was also performed. It was found that the sequence composition of signal peptides varies across the four species, especially in residues flanking the cleavage site. Another interesting result from this study was that cDNAs encoding *S. mansoni* elastases that were fused to a reporter gene did not result in surface expression, despite the fact that elastases are known to be secreted by the invasive larval stage of schistosomes (Pearson, McManus et al. 2005). It was suggested that this was either due to the transfected COS7 cells used in this method being able to recognize some, but not all, signal sequences employed by *S. mansoni,* or other incompatibilities in protein trafficking in different species that cause incorrect reporter processing. **The implication is that traditional signal peptides used for secretion of proteins by other eukaryotes may not be employed by the invasive larvae of schistosomes** (for traditional signal peptides, see Current Understanding- Protein Secretion and Signal Peptides, page 35).

The accumulation of large amounts of genomic data for *S. mansoni* has enabled recent studies of its proteome, the complete set of proteins expressed by the organism. A study of *S. mansoni* cercarial secretions, which employed 1D gel electrophoresis combined with LC-MS/MS, identified a total of 172 proteins from skin lipid-induced, mechanically induced, and uninduced samples (Knudsen, Medzihradszky *et al.* 2005). A study of the *S.*

mansoni adult tegumental sub-proteome, which also involved the use of 1D and 2D gel electrophoresis combined with LC-MS/MS and MALDI TOF-TOF, identified 740 proteins, 43 of which were found to be unique to the tegument (van Balkom, van Gestel *et al.* 2005). The first proteomic analysis of soluble proteins from four different life cycle stages of *S. mansoni* was performed using 2D gel electrophoresis and resulted in the identification of 32 proteins by MS/MS protein sequencing (Curwen, Ashton *et al.* 2004).

cDNA microarrays have been used to identify *S. mansoni* gender-associated genes (Hoffmann, Johnston et al. 2002). The first oligonucleotide microarray applied to studying *S. mansoni* was also used for studying gender-associated genes and was comprised of approximately 50% of the adult parasite's transcriptome (Fitzpatrick, Johnston et al. 2005). 197 transcripts were shown to have a "gender-biased pattern of gene expression" in adult schistosomes.

Plasmodium falciparum

P. falciparum is a protozoan (unicellular) parasite that causes malaria. Plasmodia live within the host's erythrocytes surrounded by a lipid membrane, termed the parasitophorous vacuolar membrane (Lingelbach and Joiner 1998). An ER-type signal sequence (see Background- Protein Secretion and Signal Peptides) is responsible for parasitic proteins being exported across the parasitic plasma membrane and into the lumen of the parasitophorous vacuole (Wickham 2001). It has been found that additional signaling sequences are necessary for proteins to be transported across the parasitophorous vacuole into the erythrocyte cytoplasm.

Through the use of fluorescence microscopy, two different *P. falciparum* histidine-rich proteins (histidine-rich protein I, PfHRPI, and histidine-rich protein II, PfHRPII) were found to contain a domain, immediately after the signal sequence and before a histidine-rich region, that is necessary for export of green fluorescent protein (GFP) from the lumen of the parasitophorous vacuole to the erythrocyte cytoplasm (Lopez-Estraño, Bhattacharjee *et al.* 2003). For PfHRPI this domain was SNNCNNGNGSGDSFDFRNKRTLAQKQ and for PfHRPII this was FNNNLCSKNAKGLNLNKRLLYETQAHVDD. In both cases the domain was within 40 amino acids of the signal sequence and needed to be exposed at the N-terminus in order for export to the cytoplasm to occur.

Through N-terminal sequence alignment a conserved pentameric sequence motif, RxLxE, referred to as the Pexel motif (*Plasmodium* export element), was also found at the N-terminus of other exported proteins (Marti, Good *et al.* 2004). The Pexel motif occurs in about 160 soluble and insoluble cytoplasmic-side proteins within 60 amino acids of an upstream signal sequence, and in about 225 virulence proteins, some of which lack the signal sequence. The motif was verified by fusing fluorescent signaling proteins to proteins where this pentameric motif had been either removed or mutated and demonstrating that this did not result in fluorescence being exported to the cytoplasm. An eleven amino acid motif, RxSRILAExxx (RILAE forms the core), referred to as a vacuolar transport sequence (VTS), that has features in common with the Pexel motif and also occurs 60 amino acids downstream of a signal sequence, was found in about 320 proteins through the reiterative use of MEME and MAST (Hiller, Bhattacharjee *et al.* 2004). It was experimentally proven to be responsible for both soluble and insoluble proteins being exported from the parasitophorous vacuole to the erythrocyte cytoplasm. It has recently been suggested that the region necessary for export in *P. falciparum* is actually larger (~30 aa) than the Pexel motif, but that Pexel is the core of the region and contains the most conserved amino acids (Bhattacharjee, Hiller et al. 2006). Most interestingly, it has also been found that in hundreds of secretory proteins in a divergent eukaryote, *Phytophthora,* there is conservation of the RxLR motif (Bhattacharjee, Hiller et al. 2006). *Phytophthora* is an oomycete that is the causative agent of potato late blight. Three different species have been shown to contain the RxLR motif in a position that is conserved across *Plasmodium* and *Phytophthora*. It was also found that there is conservation of E/D residues with 25 amino acids downstream of the RxLR motif in *Phytophthora* species. It is important to note that predictive algorithms used to find the plasmodial motif did not recognize the RxLR motif in *Phytophthora.* Rather this motif was experimentally proven to have equivalent function in the two species by expressing a chimera of GFP and a 50 amino acid sequence from the *Phytophthora* avirulence protein AVR3a containing the RxLR motif in *P. falciparum.* The result was expression of GFP in the cytoplasm of *P. falciparum.*

A comparison of the VTS and Pexel motifs and the methodologies used to discover them found that the pentameric Pexel motif is contained within the 11 amino acid VTS motif, but that there is a 30% difference in their corresponding sets of secreted proteins (Haldar, Hiller *et al.* 2005). It is postulated that the differences in the motifs are a result of the two different methodologies used: the VTS motif was found using the MEME position-specific probability matrix, while the Pexel motif was found through sequence alignment and using a linear two exon-structured pattern to search databases of predicted proteins. There is a *P. falciparum* protein known to be exported that does *not* contain the two exon structure. Two different secretion models are proposed (Haldar, Hiller *et al.* 2005). In the first model, proteins are secreted from the lumen of the parasitophorous vacuole to the lumen of specialized sorting and trafficking organelles, similar to the Golgi called Maurer's clefts, via vesicles and then into the erythrocyte cytosol or plasma membrane. In the second model, proteins are delivered across the plasma membrane of the parasitophorous vacuole into the erythrocyte cytoplasm, where they associate with the cytoplasmic side of the Maurer's clefts.

Through N-terminal alignment of 60 PfEMP1 proteins it was found that the virulence protein encodes a motif, linearized to SAKHLLDRIGKDVHDQVK, that is different from the motif (Pexel) found in other virulence proteins, but that shares features with it (Marti, Good *et al.* 2004). Hiller *et al.* discovered a different motif in PfEMP1 proteins, QFFRWFSEWSE, through reiterative use of MEME and MAST using an earlier discovered motif (VTS) (Hiller, Bhattacharjee *et al.* 2004).

B. Current Understanding

S. mansoni

There are 5 species of *Schistosoma* that infect humans: 4 intestinal schistosomes including *S. intercalatum, S. japonicum, S. mekongi, S. mansoni*, and 1 urinary schistosome named *S. haematobium*. Each species has a specific freshwater snail as its intermediate host. 85% of schistosomiasis cases occur in sub-Saharan Africa. *S. haematobium* is the most prevalent and widespread species in Africa and the Middle East. *S. intercalatum* occurs in 10 countries in the rainforest belt of Africa. *S. japonicum* is restricted to the Pacific region including China and the Philippines. *S. mekongi* is found in limited areas of Laos and Cambodia. *S. mansoni* is found in Africa and is the only species seen in South America including Brazil, Venezuela, and Surinam.

The life cycle of *S. mansoni* consists of 6 stages (Figure 2). Parasitic eggs contained in human feces hatch into a larva, referred to as miracidia, upon contact with fresh water. The miracidia penetrate the tissue of their snail intermediate host, *Biomphalaria glabrata*, in which they develop into sporocysts. Sporocysts multiply asexually, producing hundreds of thousands of cercariae. The cercariae are free swimming and are induced by human skin lipid to lose their tails and penetrate human skin, at which time they are transformed into schistosomula. Shistosomula use their host's circulatory system to migrate to the blood vessels of the liver, intestine, or bladder. They then sexually mature into adult schistosomes, form pairs, and live in copula for the rest of their lives, laying 300 eggs per day.

Schistosomiasis is caused by an immune response to eggs that are lodged in tissues, failing to be excreted (Capron and Capron 1994). Granulomas form in response to the eggs and result in fibrosis that can cause enlargement of the liver, urinary obstruction in the bladder, and kidney damage. Clinical signs do not appear until the worm burden increases through repeated infection causing an increase in the number of granulomas. Symptoms can include abdominal pain, cough, diarrhea, extremely high white blood cell count, fever, fatigue, and enlarged liver and spleen. People repeatedly infected can have liver, intestine, lung, and bladder damage.

The *S. mansoni* is a diploid organism with seven pairs of somatic chromosomes and one pair of sex chromosomes and a genome that is approximately 270 Mbp (Simpson, Sher *et al.* 1982), about one tenth the size of the human genome. It is at least 30% repetitive (Le Paslier, Pierce *et al.* 2000), may be as much as 60% repetitive (Johnston, Blaxter *et al.* 1999), and has a GC content of 34% (Hillyer 1974). Recent findings suggest that much of this repetitive composition is comprised of mobile genetic elements, or retrotransposons, which have likely played a large role in their evolution (Brindley and Yoshino 2003). The number of expressed genes is estimated to be 15-20,000.

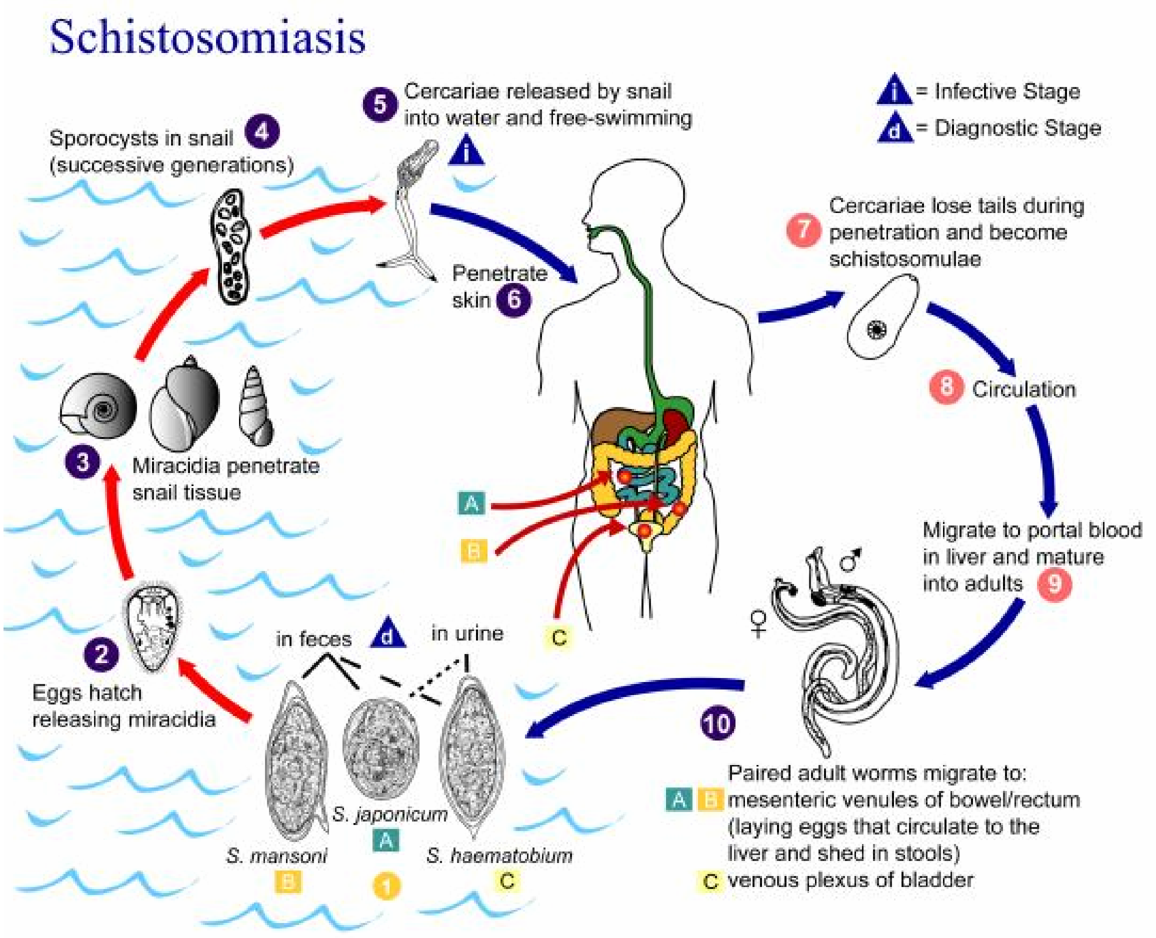

Schistosomiasis
i = Infective Stage
d = Diagnostic Stage
5 Cercariae released by snail into water and free-swimming
i
4 Sporocysts in snail (successive generations)
Penetrate skin 6
7 Cercariae lose tails during penetration and become schistosomulae
8 Circulation
3 Miracidia penetrate snail tissue
A
B
C
Migrate to portal blood in liver and mature into adults 9
♂
♀
2 Eggs hatch releasing miracidia
in feces
d
in urine
10
Paired adult worms migrate to:
A B mesenteric venules of bowel/rectum (laying eggs that circulate to the liver and shed in stools)
C venous plexus of bladder
S. japonicum
A
S. mansoni
B
1
S. haematobium
C

Figure 2. Life cycle of schistosomes
Center for Disease Control's Public Health Image Library ID #3417
Content Provider: CDC/Alexander J. da Silva, PhD/Melanie Moser
Copyright restrictions: None (http://phil.cdc.gov/Phil/home.asp)

The chromosomes of *S. mansoni* condense during cell division, making it possible to view them via light microscopy (Johnston, Blaxter *et al.* 1999). They are large with varying sizes, shapes, centromere positions, and banding patterns. Individual chromosomes cannot be isolated because even the smallest chromosome is too large to be separated using pulsed field gel electrophoresis (PFGE) (Johnston, Blaxter *et al.* 1999), but chromosome maps can be generated through such techniques as fluorescent in situ hybridization (FISH) and primed in situ hybridization (PRINS) techniques (Short, Liberatos *et al.* 1989).

Two of the three sample sets used in this study involve proteins released from the *S. mansoni* cercarial stage just before it penetrates the skin and infects its human host. The ventral sucker (acetabulum) plays an active role in skin exploration and penetration. During skin exploration it deposits mucus and attaches to these deposits alternately with the oral sucker. It serves to anchor the parasite to the skin while the oral sucker works its way through layers of skin (Stirewalt and Kruidenier 1961).

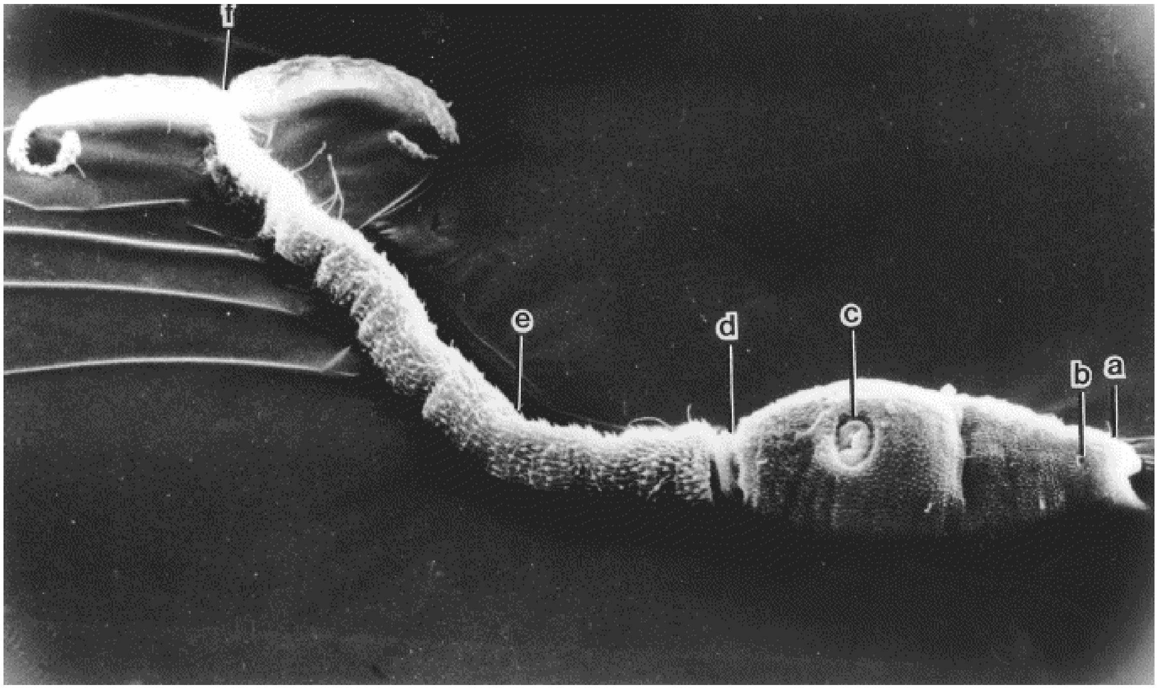


Figure 3. Scanning electron micrograph of a *S. mansoni* cercariae
(a) oral sucker, (b) mouth, (c) acetabulum/ventral sucker,
(d) body/tail junction, (e) tail and (f) tail bifurcation.
Spines are easily visible covering both the body segment and tail.
The entire length of the organism is approximately 500 µm,
but can vary considerably due to its ability to contract and elongate.
Reprinted with permission from Dr. Fred Lewis, Dr. Carolyn Cousin
and the publishers of Micron. (Dorsey, Cousin et al. 2002).

Schistosome cercaria are reported to have no fewer than four different types of glands including 1) the escape glands, 2) the head gland, and 3) two sets of acetabular glands-pre and post (Schmidt and Roberts 2000). The escape glands have been shown to expel their contents when cercaria emerge from their snail host (Schmidt and Roberts 2000). The head gland has narrow duct openings into the tegument in cercariae that have recently emerged from their snail hosts. In contrast, in new schistosomules the head gland has greatly expanded ducts which are filled with secretory bodies (Dorsey, Cousin et al. 2002), which likely function in the adjustment of the schistosomulae post-penetration. The acetabular glands are named according to their position relative to the acetabulum (ventral sucker). Two pairs of preacetabular glands are located anterior to the acetabulum and three pairs of postacetabular glands are located posterior to the acetabulum. The preacetabular glands produce mucus and help cercariae adhere to surfaces, while the postacetabular gland vesicle secretions contain various enzymes, including proteases, as well as high levels of calcium, making them important in skin penetration. It has been shown that an influx of calcium is necessary but not sufficient for release of proteinases from cercarial preacetabular glands (Fusco, Salafsky et al. 1991). Figures 4 and 5 provide two different perspectives of the head and acetabular glands.

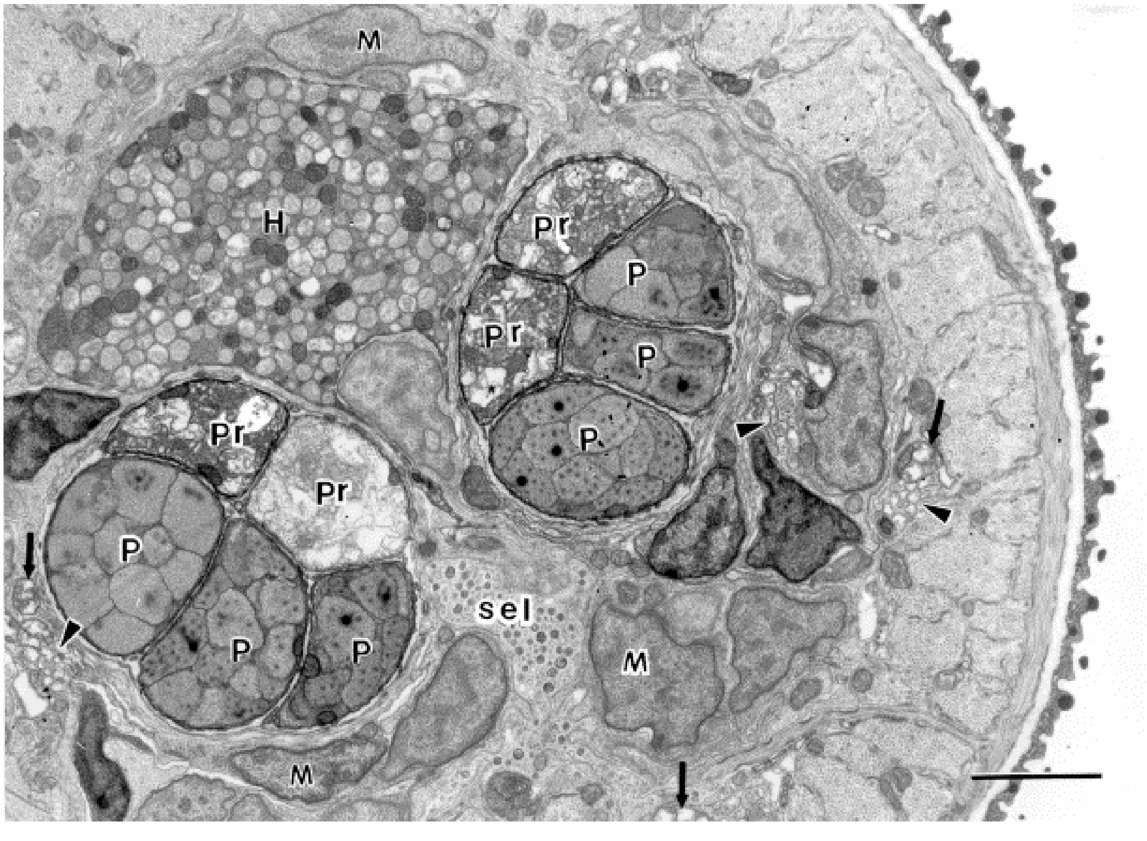


Figure 4. Cross section of a *S. mansoni* cercaria
Ventral area of the anterior organ near the esophagus.
Head gland (H), two bundles each containing two preacetabular
glands (PR), three postacetabular glands (P). Scale bar 1.5 µm.
Reprinted with permission from Dr. Fred Lewis, Dr. Carolyn Cousin,
and the publishers of Micron (Dorsey, Cousin et al. 2002).

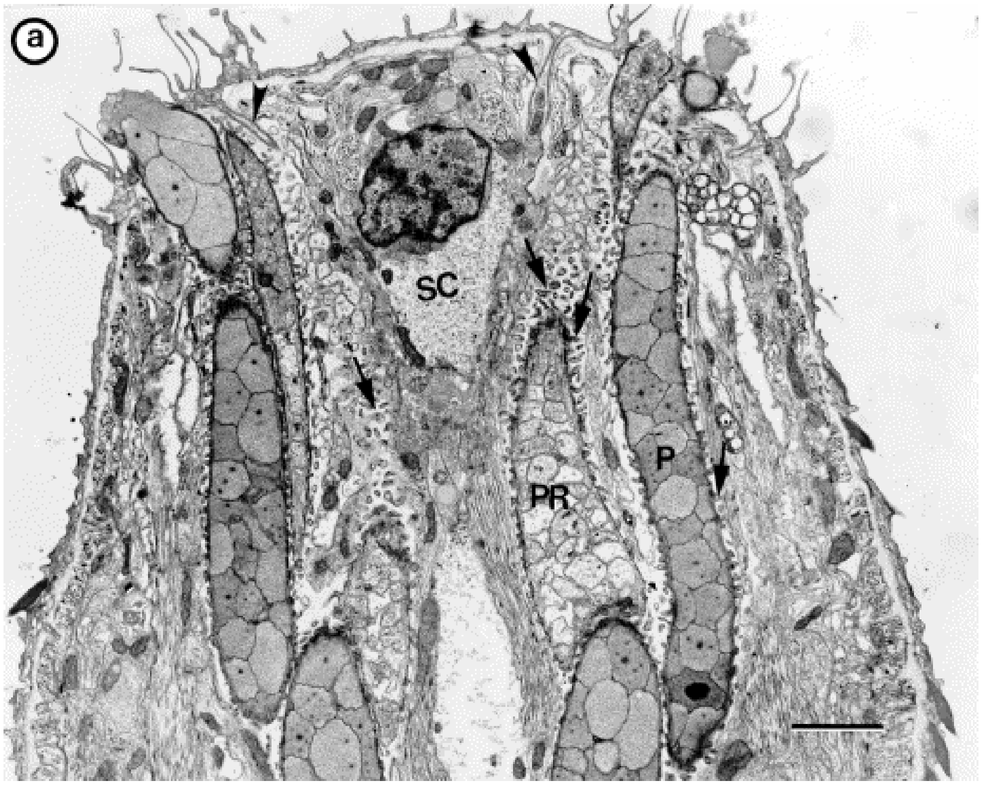


Figure 5. Anterior organ of a *S. mansoni* cercaria
A low magnification section of part of the anterior organ of a cercaria.
Shown are ducts of the pre- (PR) and postacetabular (P) glands, one
of the two anterior organ support cells (SC),head gland ducts (arrowheads),
and microtrichs (cytoplasmic processes) projecting from the limiting
membrane of the acetabular gland ducts (arrows). Scale bar 1.5 µm.
Reprinted with permission from Dr. Fred Lewis, Dr. Carolyn Cousin,
and the publishers of Micron (Dorsey, Cousin et al. 2002).

Schistosomes are eukaryotic organisms of the Kingdom Metazoa, Phylum Platyhelminthes, Class Trematoda, Subclass Diegenea, Order Strigeidida, Family Schistosomatidae, and Genus Schistosoma. They are often referred to as blood flukes. They are a unique family among the Digeneans, along with two other families, Spirorchiidae and Sanguinicolidae, in that they have two host life cycles, a snail intermediate host and a vertebrate (human) definitive host; most Digeneans have three host life cycles including a metacercariae stage (Loker and Mkoji 2005). The tegument of all three of these families is also unique in that it is bound by a double lipid bilayer, whereas other flukes have only a single lipid bilayer (McLaren and Hockley 1977). This additional lipid bilayer is most likely needed because the adult schistosome lives in the vascular system of its definitive human host.

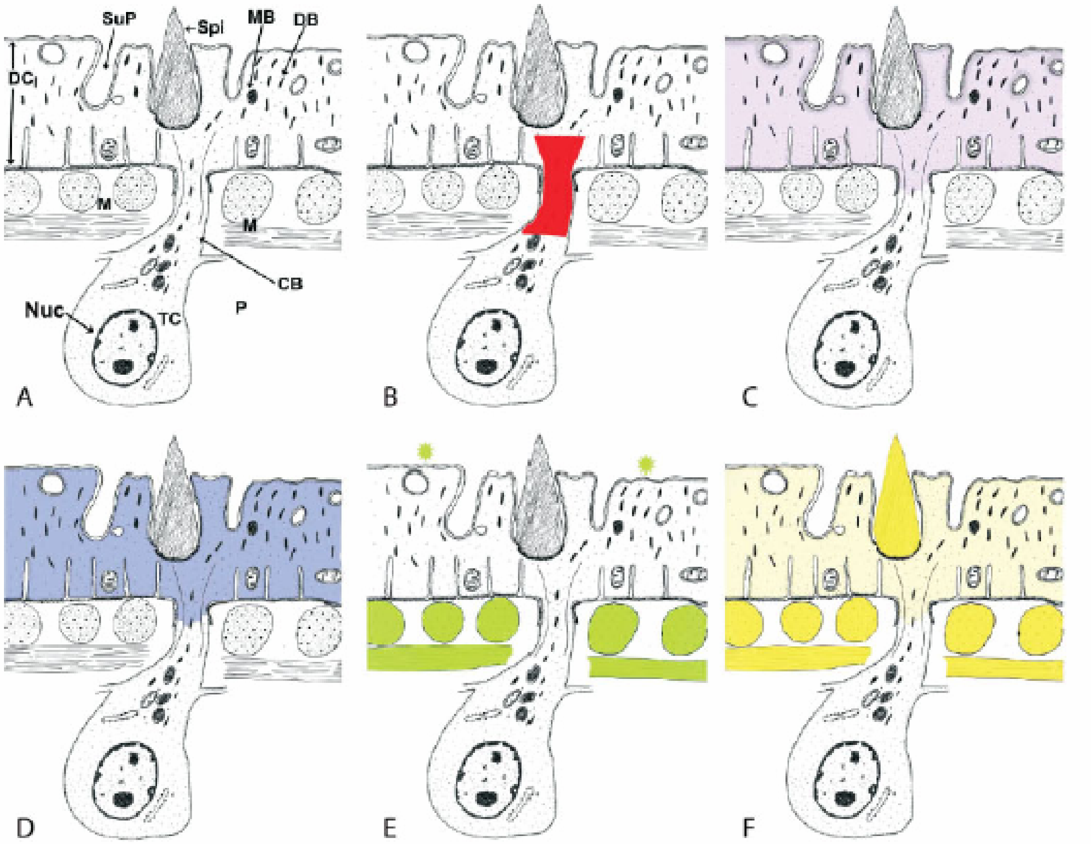


Figure 2. Structure of the schistosome tegument and distribution of key cytoskeletal and motor components. This illustration is based on ultrastructural models of the dorsal surface of the tegument of an adult male *S. mansoni*.[15] **A:** General features. **B–F:** Distribution of: **B:** Microtubules (red). **C:** Dynein light chains (purple). **D:** Tegument-associated proteins, a family of calcium-binding molecules with sequence identity with dynein light chains (blue). **E:** Paramyosin (and myosin) (green). Asterisks at surface indicate the observation of paramyosin-like immunoreactivity at the surface of the tegument in schistosomula; **F:** Actin (yellow). Abbreviations: CB, cytoplasmic bridges; DB, discoid bodies; DC, distal cytoplasm; M, myofibrils; MB, membranous bodies; Nuc, nucleus of tegumentary cyton; P, parenchyma; SuP, surface pits; Spi, spine; TC, tegumentary cytons.

Figure 6. Components of the *S. mansoni* tegument
Reprinted with permission from Dr. Malcolm Jones
(Jones, Gobert et al. 2004).

The morphology of the tegument and its involvement in host-parasite interactions has been extensively studied. The tegument undergoes changes as the schistosome passes through the different stages of its life cycle. Miracidia possess a tegument consisting of a layer of ciliated sheets, which is shed just before they infect their snail intermediate host. Cercaria possess a tegument that has an osmotic-protective, but highly immunogenic glycocalyx layer. This is shed within a few hours of penetrating the human host and is replaced by an immune-evasive adult tegument (Jones, Gobert et al. 2004). Most characteristics of the tegument remain consistent throughout the different life cycles.

Proteins contained in postacetabular secretions (sample sets 1 and 2) are important to study because they are directly involved in initial human infection; they contain enzymes involved in breaking down skin components that allow the parasite to find its way into the bloodstream. Proteins contained in the tegument (sample set 3) are also important to study because the tegument serves as the major interface between the schistosome and its environment. Proteins isolated from postacetabular vesicles or the tegument are likely to contain antigenic proteins that could be good targets for drug or vaccine development.

Protein Secretion, Signal Peptides, and Propeptides

Proteins often start out as inactive pre-pro-peptide precursors. The pre-domain consists of a signal peptide which directs the protein to the appropriate cellular compartment during synthesis and gets cleaved during this translocation process, while the pro-domain can have various functions once it is cleaved such as protein localization, regulation of activity, and involvement in protein folding. In secreted proteins the pre-domain is the signal that causes transport into the endoplasmic reticulum (ER) and, subsequently, the trans Golgi network (TGN), and the pro-domain is the signal that causes the proteins to be directed into the regulated secretory pathway for eventual export out of the cell (Alberts, Bray et al. 1994).

Signal peptides that direct proteins to the lumen of the ER are referred to as ER-type signal peptides. Signal peptides consist of a positively charged N-terminal region, a hydrophobic central segment, and a polar, but neutral, C-terminal region. Their structure is said to follow the (-3, -1) rule which states that in order for cleavage to occur correctly, small and neutral amino acids must exist at positions -3 and -1 relative to the cleavage site (Von Heijne 1983). ER-type signal peptides are cleaved by signal peptidase I. N-terminal anchor sequences are also capable of delivering peptides to the ER, of which there are several types. If only a signal peptide is present, the ribosome synthesizing the protein is signaled to deliver it to the lumen of the ER (High, Flint et al. 1991), while if either both a signal peptide and a signal anchor sequence is present or a signal anchor sequence alone is present, the protein is anchored into the ER membrane instead (Wahlberg and Spiess 1997).

All eukaryotic cells have an endoplasmic reticulum (ER), which plays a central role in lipid and protein biosynthesis. The membrane of the ER is the site of synthesis of the transmembrane proteins for most of the cell's organelles and all proteins that will be secreted outside the cell are initially delivered to the lumen of the ER. The ER is made up of a convoluted, interconnected set of tubules and sacs that extend throughout the cytosol and make up a continuous space called the ER lumen.

Once in the lumen of the ER, proteins are glycosylated, molecular chaperones aid protein folding, and misfolded proteins are identified and retrotranslocated to the cytosol, where they are degraded by a proteasome. Properly-folded proteins enter the Golgi apparatus, where the glycosylation of the proteins is modified and further posttranslational modifications, including cleavage and functionalization, may occur. They are then translocated to the cell surface via transport vesicles, in which further protein modification may occur. This is referred to as the constitutive secretory pathway and it operates in all cells. All eukaryotic cells require this type of secretion where proteins and small molecules leave the trans Golgi network in a steady stream in transport vesicles. These vesicles fuse with the plasma membrane causing their soluble contents to be released to the extracellular space, while the insoluble components provide new components for the cell's plasma membrane as they become incorporated into it.

Some of the proteins contained in the ER may contain a signal, or pro-domain, that causes them to either be retained in the ER or shuttled off into special secretory vesicles. Secretion via secretory vesicles is referred to as the *regulated* secretory pathway. It operates only in specialized secretory cells and is used to secrete soluble proteins and other substances such as hormones and enzymes rapidly and on demand. Similarly to transport vesicles, secretory vesicles fuse with the plasma membrane causing their soluble contents to be released to the extracellular space, but not until an extracellular signal is received (Alberts, Bray et al. 1994). Proteins destined for secretory vesicles accumulate in the trans Golgi network (TGN). The mechanism behind subsequent uptake into secretory vesicles is not fully understood, but is thought to resemble phagocytosis. When secretory vesicles bud from the TGN, their contents become greatly condensed.

As in the non-regulated secretory pathway, many of the proteins contained in secretory vesicles and destined to be exported are thought to undergo one or more rounds of cleavage, either while still in the TGN, in the secretory vesicles, or in the extracellular fluid after secretion has occurred.

Pro-domains become exposed at their N-terminal once the signal peptide is cleaved. Pro-domains are often cleaved immediately C-terminally of motifs containing multiple basic residues (Duckert, Brunak et al. 2004). A family of evolutionarily conserved dibasic-specific and monobasic-specific Ca^{2+}-dependent serine proteases called subtilisin/kexin-like proprotein convertases (PCs) are responsible for this type of cleavage. In mammals PCs are the major endoproteolytic processing enzymes in the secretory pathway and are involved in processing an extensive number of proteins including peptide hormones, receptors, extracellular matrix proteins, and glycoproteins from infectious viruses and bacterial toxins (Duckert, Brunak et al. 2004). They usually cleave immediately after the consensus sequence [R/K]-X_n-[R/K] where R represents arginine, K represent lysine, X represents any amino acid, and n represents 0, 2, 4, or 6. One of the aims of this thesis study is to reveal pro-domains recognized by PCs or other proteolytic enzymes in proteins secreted via acetabular secretory vesicles in schistosomes.

Signal Peptide Prediction

Original methods for predicting signal peptides and their cleavage sites used weight matrices (Heijne 1986). More recent methods use neural networks. SignalP, which uses neural networks, was used in this thesis study as it has shown accuracy, specificity, and sensitivity over other non-weight matrix type methods (Bendtsen, Nielsen et al. 2004) including PSORT-II (Nakai and Horton 1999) and SubLoc (Hua and Sun 2001). It uses one neural network to predict signal peptides and another one to predict signal peptidase I cleavage sites. (LipoP predicts signal peptidase II cleavage sites, which are found in lipoproteins (Juncker, Willenbrock et al. 2003).)

In the first version of SignalP (Nielsen, Engelbrecht et al. 1997) data for training the eukaryotic neural networks was obtained from SWISS-PROT version 29 (Bairoch and Boeckmann 1994) and included nonredundant secreted (first 30 amino acids), membrane, cytosolic and nuclear proteins (first 70 amino acids). A total of 1011 proteins from eukaryotes and an additional 416 specifically from humans were used. The network was a feed-forward network, trained using back-propagation. Test performance was determined by using 5-fold cross-validation (data sets were divided in five approximately equal parts and each part was used as a test set while the other 4 parts were used to train the network) and taking an average over each of the five test sets. Correlation coefficients were calculated based on the number of correctly and incorrectly predicted signal peptides. It was found that for the eukaryotic sample set, the correlation coefficient for signal peptide discrimination was .97 and the percent of correctly predicted cleavage sites was 70.2%. Eukaryotic signal peptides were shown to be dominated by Leu with some occurrence of Val, Ala, Phe, and Ile. The subset of human proteins showed no significant difference from the eukaryotic data set in their signal peptides and training the networks on a single species did not improve predictive performance. While the first version of SignalP was quite good at distinguishing between signal peptides and non-signal peptides, it had an error rate of about 5/% in distinguishing between signal peptides and signal anchor sequences, which are often quite similar to signal peptides after their hydrophobic region.

SignalP version 3.0 was used in this thesis study. Improved performance in this version over previous versions of the neural network algorithm was obtained by manually correcting annotation errors contained in training data (1192 eukaryotic sequences from SWISS-PROT), taking into account novel amino acid frequencies and positions for detecting annotation errors that were possibly not removed from training sequences, and implementing a new D-score for classification of signal peptides versus non-signal peptides (Bendtsen, Nielsen et al. 2004). Prediction of signal peptides *versus* non-signal peptides improved from a correlation coefficient of 97 to 98, and prediction of signal peptide cleavage sites improved from 70.2% to 79%.

Propeptide/Conserved Motif Prediction

Motif identification tools have traditionally taken one of two approaches- either alignment methods, which are *sequence-driven*, or enumerative/exhaustive methods which are *pattern-driven* (Ohler and Niemann 2001). Tools that take the alignment approach include CONSENSUS, MEME, and BioProspector. CONSENSUS (Hertz and Stormo 1999) performs a local multiple sequence alignment of all sequences, aligning them one at a time and constructing a weight matrix that optimizes the information content. MEME (Bailey and Elkan 1994), which applies expectation maximization, and BioProspector (Liu, Brutlag et al. 2001), which uses a Gibbs sampling strategy, consider start positions of the motif to be unknown and determine which positions give the most conserved motif by performing a local optimization (Ohler and Niemann 2001). Tools that take the enumerative/exhaustive approach update a nucleotide probability matrix as they evaluate the frequency of occurrence of all possible sequences of length *n* using an iterative process. A new approach that claims advantages over existing algorithms employs evolutionary computation (Fogel, Weekes et al. 2004).

MEME, a sequence-driven alignment method, was used in this thesis study. It uses an unsupervised learning approach to searching for motifs in data sets about which very little is known. The only inputs that MEME requires are the set of sequences and the width of motif to search for. It can find several different motifs with differing numbers of occurrences in a single dataset. An algorithm called MM, mixture model, is used to fit a two-component finite mixture model to the set of sequences (Bailey and Elkan 1994). The two components consist of: 1) probability densities of the motif itself and 2) probability densities for all other positions in the sequences, the background. The expectation maximization technique (Aitkin and Rubin 1985) is then applied to the two densities to estimate each of their parameters and the mixing parameter.

A log likelihood ratio is calculated for each column in the motif:

For I = 1…..n, where n = number of columns in the motif, llr = ∑ log (P(site-specific occurrence|motif)/P(occurrence|background model))

The first probability in this equation is one of the outputs of MEME, a position-specific probability matrix, or PSPM, which shows probabilities of each amino acid vs. every position in a particular motif. It is based on the set of input sequences. The second probability in the equation is not position-specific. It is based on either the input sequences or any other set of specified sequences.

MEME computes a p-value for the log likelihood ratio (log L) of each column in the motif, instead of a p-value of the sum of the log likelihood ratios, and then arrives at an e-value for the motif by computing the product of the p-values. This makes the e-value reported by MEME an approximation, which is much more efficient to compute, rather than an actual e-value. The e-value corresponds to the number of motifs that would have an equal or higher log likelihood ratio if the training set of sequences had been generated randomly according to the background model. The e-value is also based on the width of the motif, the number of occurrences, the size of the training set, and the type of distribution. MEME outputs the motif with the smallest e-value, represented as a position-specific scoring matrix, PSSM.

Given the missing data Z, the observed data X, and the current parameter values, $\theta = \theta^{(0)}$ and $\lambda = \lambda^{(0)}$, the e-value for the first component of the mixture model (the motif itself), is computed as:

$$\underset{(Z|X,\theta^{(0)},\lambda^{(0)})}{\mathrm{E}}[\log L(\theta,\lambda|X,Z)] \quad = \quad \underset{(Z|X,\theta^{(0)},\lambda^{(0)})}{\mathrm{E}}[\sum_{i=1}^{n}\sum_{j=1}^{g} Z_{ij}\log(p(X_i|\theta_j)\lambda_j)]$$

The e-value for the second component of the mixture model (all other positions in the sequences) is computed as:

$$Z_{ij}^{(0)} \quad = \quad \mathrm{E}[Z_{ij}|X,\theta^{(0)},\lambda^{(0)}]$$

High log likelihood ratios correspond to low the e-values. Generally the e-value for the motif should be lower then .01 to be considered real.

The total information contained in each motif can be determined by summing up the information contained in each position of the motif. Positions that are highly conserved have high information content (IC, measured in bits), while positions that are not highly conserved have low information content. MEME documentation states that for a motif to be useful for database searches, it must contain at least $\log_2(N)$ bits of information, where N is the number of sequences in the database being searched. For example, when searching a database of 20,000 sequences for a MEME motif, the IC of the motif should be at least 14.4.

MAST (Bailey and Gribskov 1998) uses the position-specific scoring matrix and threshold produced by MEME together as a Bayes-optimal classifier to search for occurrences of the motif in databases. A match score is computed for each sequence in the specified database. The score for the match of a position in a sequence to a motif is computed by summing the appropriate entry from each column of the position-dependent scoring matrix that represents the motif. The QFAST algorithm (Bailey and Gribskov 1998) is then used to compute the combined p-value of a sequence, which measures the strength of the match of the sequence to all the motifs. It is calculated by 1) finding the score of the single best match of each motif to the sequence, 2) calculating the sequence p-value of each score (defined as the probability of a random sequence of the same length containing some match with a score as good or better), 3) forming the product of the p-values, and 4) taking the p-value of the product. The E-value of a sequence is the expected number of sequences in a random database of the same size that would match the motifs as well as the sequence does. It is equal to the combined p-value of the sequence times the number of sequences in the database. A motif occurrence is defined as a position in the sequence whose match to the motif has position p-value less than 0.0001. The position p-value of a match is the probability of a single random subsequence of the length of the motif scoring at least as well as the observed match.

Protein Domain and Gene Ontology Analysis

Assignment of a particular functional domain and Gene Ontology (GO) terms to a protein can help determine potential relationships to other proteins as well as physiochemical properties. There are multiple and various tools for searching protein fingerprint databases. InterProScan (Mulder, Apweiler et al. 2005) is a tool that unites all protein signature recognition methods and databases of all of its consortium members, making it the most comprehensive tool available. These member databases and methods include ProDom/BlastProDom (Blastall), PRINTS/FingerPrintScan, TIGRFAMs/Hmmpfam, Pfam/Hmmpfam, PROSITE/ScanRegExp + ProfileScan, PIRSF/Hmmpfam, CATH/Hmmpfam, PANTHER/Hmmsearch, SignalPHMM/SignalPHMM, Transmembrane/TMHMM2.0, SMART/Hmmpfam, and SUPERFAMILY/Hmmpfam. The last four listed are not public and so will not be used. InterProScan launches all of these various applications, taking advantage of pre-computed results whenever possible, and then merges all the data together. GO annotation data can also be included in the output of InterProScan and included in the merged result file.

C. Research Question

Are there any significant common motifs among secreted and/or tegumental proteins and, if so, what role do they play in *S. mansoni* secretion?

D. Intended Research Project

The intent of this project is to follow the same workflow that Hiller *et al.* (2004) used to find an export signal in *P. falciparum* sequences. MEME will be used to perform motif analysis on cercarial secretion and adult tegument proteins from *S. mansoni* to determine if a conserved signal is present. If such a signal is found, MAST will be used to search for the motif in public *S. mansoni* databases. Recursive iterations of MEME and MAST will be used to continually optimize the motif and search for it in public databases until the motif cannot be refined further and it is not revealed in any new sequences. Bioinformatic analysis of the motif and the proteins found to contain it will be used to propose a secretion model.

III. METHODS

A. Materials: Software and Hardware

NCBI's BLAST 2.2.1 (Altschul, Madden et al. 1997) was executed on a UNIX server (6-337 MHz processors, 4 Gb RAM, 600 Gb disk space) for performing local sequence alignments to determine sequence similarity. ClustalW (Thompson, Higgins et al. 1994) was executed from the European Bioinformatics Institute's web server (http://www.ebi.ac.uk/clustalw/) for performing multiple sequence alignments and generating phylogenetic trees, and SignalP 3.0 (Bendtsen, Nielsen et al. 2004) was executed from the SignalP web server (http://www.cbs.dtu.dk/services/SignalP/) for predicting signal peptides. The following software was installed and executed from a UNIX server (2 Xeon 3.5 MHz processors, 4 Gb memory, 1.8 Tb disk space): MEME, Multiple Expectation Maximization for Motif Elicitation, (Bailey and Elkan 1994) version 3. 0 for predicting conserved sequence motifs, MAST, Motif Alignment Search Tool, (Bailey and Gribskov 1998) version 3.0 for searching databases for motifs predicted by MEME, and InterProScan version 12.0 (Mulder, Apweiler et al. 2005) for protein domain and Gene Ontology (GO) annotation. The same machine was used to run Matlab 7.0.1 (The Mathworks) for developing a script to compute the entropy of sequence protein domain and GO features. Perl 5.8.4, including the BioPerl module, was used to develop scripts that automate the parsing and reformatting various result files.

B. Samples and Subjects

1. Animal or Human Subject Clearance

This project did not involve any wet lab work; there was no direct contact with the parasite of interest.

2. Sample Size

A total of 177 proteins were analyzed. This included two different sample sets from *S. mansoni* cercarial secretions (Knudsen, Medzihradszky et al. 2005) and one from *S. mansoni* adult tegument (van Balkom, van Gestel et al. 2005) as follows: 1) 81 *S. mansoni* cercarial secretion proteins induced by skin lipid (average length = 423 aa), 2) 53 *S. mansoni* cercarial secretion proteins induced by tail shearing (average length = 347), 3) 43 adult tegument proteins (average length = 224) and 4) 155 full-length *S. mansoni* proteins downloaded from NCBI for use as a control set (average length = 434). To distinguish between the two different sets of cercarial proteins, the 81 proteins from tail shearing induction will be referred to as "vesicles" and the 53 proteins from lipid induction will be referred to as "secretions" hereafter. The vesicles sample set provided the most information as it contains the most sequences, has an average length of sequence that is significantly longer than the other two sample sets, and is the result of a method that minimized contamination, as explained below.

Sequence databases were downloaded for use with the MAST program as follows:

1. 455 full-length *S. mansoni* proteins NCBI's nonredundant protein database (accessed July 17, 2006)
2. 3894 full-length *S. mansoni* and *S. japonicum* proteins from NCBI's nonredundant protein database (accessed April 23, 2005)
3. A set of 26,228 assembled EST contigs and singletons from the Schistosoma Genome Network's SmAE database (Verjovski-Almeida, DeMarco et al. 2003) (accessed September 12, 2005, url: http://cancer.lbi.ic.unicamp.br/schisto6/),

4. A set of 33,704 assembled EST contigs and singletons from TIGR's SmGI release 5 (accessed January 22, 2005), *S. mansoni* Gene Index database (Quackenbush, Liang et al. 2000) (url: http://www.tigr.org/tigr-scripts/tgi/T_reports.cgi?species=s_mansoni).

3. Samples

Three different sets of samples were used in this study as listed in Appendix A. Experiments for stimulating, isolating, and identifying the samples were completed previously by Knudsen *et al.* (2005) and van Balkom *et al.* (2005). Those experimental procedures are briefly described here for the purposes of understanding the source of the samples and methods used to induce secretion.

Two sets of samples from *S. mansoni* cercariae were isolated and identified by Knudsen *et al.* (2005). In summary, the first set of samples was collected from cercariae after vesicle secretion was stimulated by mechanical tail shearing, artificially transforming the cercariae into schistosomula. After cercariae were pushed through a small syringe, which forced their tails to be sheared, they were allowed to secrete for 1.5-2.0 hours on a Petri dish containing culture medium, medium was removed from the plate, and vesicles that were left behind on the Petri dish were collected. This method resulted in a set of samples that had minimal contamination with snail, lettuce, and human proteins from the lab environment. The second set consists of samples isolated after cercariae were exposed to human skin lipid, the parasite's natural biological stimulus, on a Petri dish for 1.5-2.0 hours in warm water. Their secretions were separated from the cercarial bodies and tails through centrifugation. The proteins in both the first and second sample sets were isolated and sequenced using 1D SDS PAGE, in-gel digestion, and LC-MS/MS. Data was analyzed using Analyst QS software (Applied Biosciences) with Mascot (Matrix Science, London, UK) and then identified using Mascot server version 2.0.01 to search both the nonredundant protein database at NCBI (September 8, 2004) and a six-frame translation of TIGR's SMGI database, version 5. These sample sets will be referred to hereafter as "vesicles" and "secretions", respectively.

The third set of samples was isolated from the adult schistosome tegument as described by van Balkom et al. (van Balkom, van Gestel et al. 2005). Briefly, the tegument surface membranes of adult worms was "stripped" through a series of steps that included washing, freezing, applying vortex pulses, passing the tegument membrane-containing supernatant over a fine stainless steel mesh, and centrifuging the filtrate to collect the tegument membrane. Proteins contained in the tegument membranes were then isolated and sequenced using both 1D and 2D SDS PAGE, in-gel digestion, immunoblotting, LC-MS/MS, and MALDI TOF-TOF (applied to proteins excised from 2D gels only). They were then analyzed using ProteinLynx Browser version 2.1 software (Micromass), and identified by searching a six-frame translation of TIGR's *S. mansoni* Gene Index database version 5 with Mascot software (Matrix Science, London, UK). Proteins that could not be identified using TIGR's database were identified using the *S. mansoni* GeneDB where possible (El-Sayed, Bartholomeu et al. 2004). This sample set will be referred to hereafter as "tegument".

A set of control proteins were derived by searching NCBI's nonredundant database for full-length *S. mansoni* cDNA sequences, downloading the resulting 686 sequences, translating them to protein, and using them as both the query and subject sequences in a BLAST search. BLAST results determined that 531 sequences were either identical or highly homologous to other sequences. These were removed from the data set, resulting in 155 nonredundant proteins that were used as a control set.

C. Procedures and Interventions

As outlined in Figure 7, Workflow #1, sequences were analyzed for the presence of ER-type signal peptides using SignalP and propeptides using MEME and MAST. Initial sequences were obtained by searching either NCBI's nonredundant protein database or TIGR's SmGI database for accession number provided in the two papers "Proteomic Analysis of *Schistosoma mansoni* Cercarial Secretions" (Knudsen, Medzihradszky et al. 2005) and "Mass Spectrometric Analysis of the *Schistosoma mansoni* Tegumental Sub-proteome" (van Balkom, van Gestel et al. 2005). TIGR's database provided nucleotide sequences and several different corresponding translations including an ESTScan translation (Iseli, Jongeneel et al. 1999), FrameFinder translation (Slater 1996-1999) and a six-frame translation. The best translation (i.e. best alignment between query sequence and subject sequence) was determined using a manual BLAST (Altschul, Madden et al. 1997) search against NCBI's nonredundant protein database. In some instances protein translations were not provided by TIGR, in which case a Perl script was used to translate the DNA sequence into six frames and the longest frame was selected and verified using BLAST.

Sequences were analyzed using SignalP version 3.0 (Bendtsen, Nielsen et al. 2004) to determine if they contain an N-terminal signal sequence. Presumably these are the sequences targeted for export via a traditional extracellular secretion model. SignalP will be set to run both its neural network trained on eukaryotic sequences and input sequences will be trimmed to 100 amino acids.

Proteins from each sample set were submitted to MEME (Bailey and Elkan 1994) to search for conserved motifs. Sequences containing a predicted signal peptide were submitted separately from those without one. Proteins with a predicted signal sequence were trimmed to 100 bases from the cleavage site predicted by SignalP before submission to MEME. If MEME found a statistically significant motif, the resulting motif file (PSSM) was used by the MAST program (Bailey and Gribskov 1998) to search three databases for the motif: 1) full-length *S. mansoni* and *S. japonicum* protein sequences from NCBI's nonredundant protein database, 2) SmGI (Quackenbush, Liang et al. 2000), and 3) SmAE (Verjovski-Almeida, DeMarco et al. 2003) for the same motif. If the motif was found in additional sequences, those sequences were combined with the ones that MEME originally found to contain the motif, and all sequences were run through MEME again to refine the initial motif. Recursive iterations of submitting sequences to MEME and motifs to MAST were performed until the motif could not be refined any further and was not found in any additional sequences. To validate these methods, identical materials and processes were followed as those employed by Hiller *et al.* to elucidate a conserved motif in *P. falciparum* and the end results were compared.

A complimentary workflow was also executed, as outlined in Figure 8, Workflow #2. To determine the amount of overlap of proteins between data sets as well as within a single data set (to determine diversity), sequence alignment methods were applied. BLAST (Altschul, Madden et al. 1997) was used to perform local alignments and CLUSTALW (Thompson, Higgins et al. 1994) was used to perform multiple sequence alignments. The alignment file resulting from the CLUSTALW alignment of each sample set was used to generate a phylogenetic tree, which was visualized using JalView (Clamp, Cuff et al. 2004). InterProScan (Mulder, Apweiler et al. 2005) was used to predict protein domains and GO terms for the *S. mansoni* cercarial secretome and adult tegument proteins. Scripts were written to process the InterProScan raw data to generate counts and percentages of protein domains and GO terms represented in each sample set, as well matrices of sequences *versus* protein domains and sequences *versus* GO terms. A Matlab script was written to calculate entropy and information gain for each feature within a dataset to determine the most significant protein domains and GO terms. These methods provided means for comparing and contrasting proteins enriched in the different sample sets for improved understanding of the set of functions used by the invading parasite.

D. Statistical Analysis

Expectation and probability scores resulting from MEME and MAST were used to determine the validity of motifs found. A Matlab script was used to calculate the entropy of protein domain and GO annotation data, from which the most significant features in each data set were determined.

E. Expected Results

The expectation is that iterative use of MEME and MAST will elucidate a common motif across multiple *S. mansoni* cercarial secretion and/or adult tegumental proteins downstream of an ER-type signal sequence. This motif will be used to infer a proposed model for the mechanism of secretion in schistosomes.

F. Alternate Plans

If reiterative use of MEME and MAST does not reveal a conserved motif, alternative methods for finding conserved motifs will be applied, such as multiple sequence alignment, ProP (Duckert, Brunak et al. 2004), and alignment using sequence logos. Functional annotation methods will also be further improved and automated.

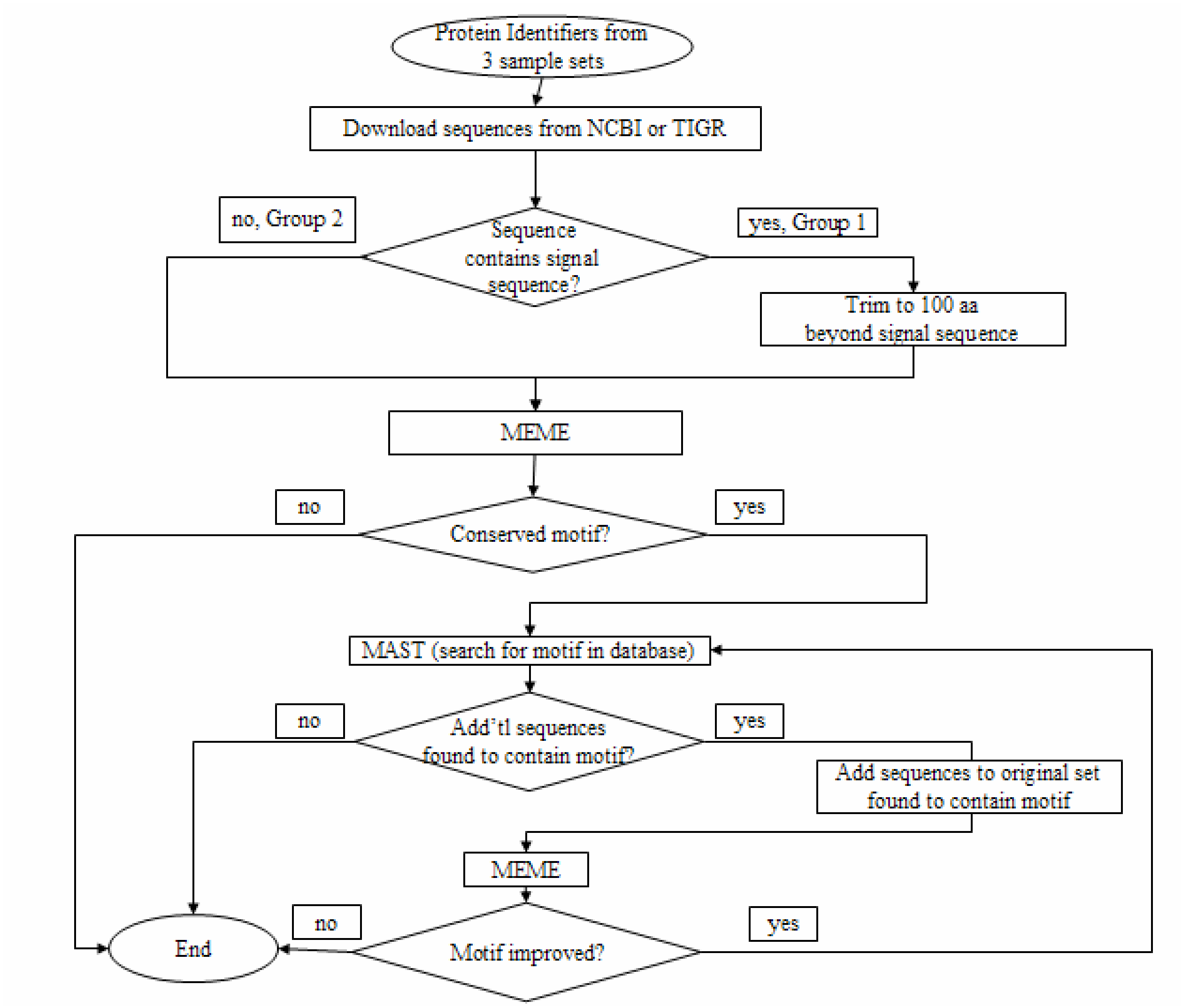

Protein Identifiers from 3 sample sets
Download sequences from NCBI or TIGR
no, Group 2
Sequence contains signal sequence?
yes, Group 1
Trim to 100 aa beyond signal sequence
MEME
no
Conserved motif?
yes
MAST (search for motif in database)
no
Add'tl sequences found to contain motif?
yes
Add sequences to original set found to contain motif
MEME
no
End
Motif improved?
yes

Figure 7. Workflow #1
Signal peptide and propeptide analysis

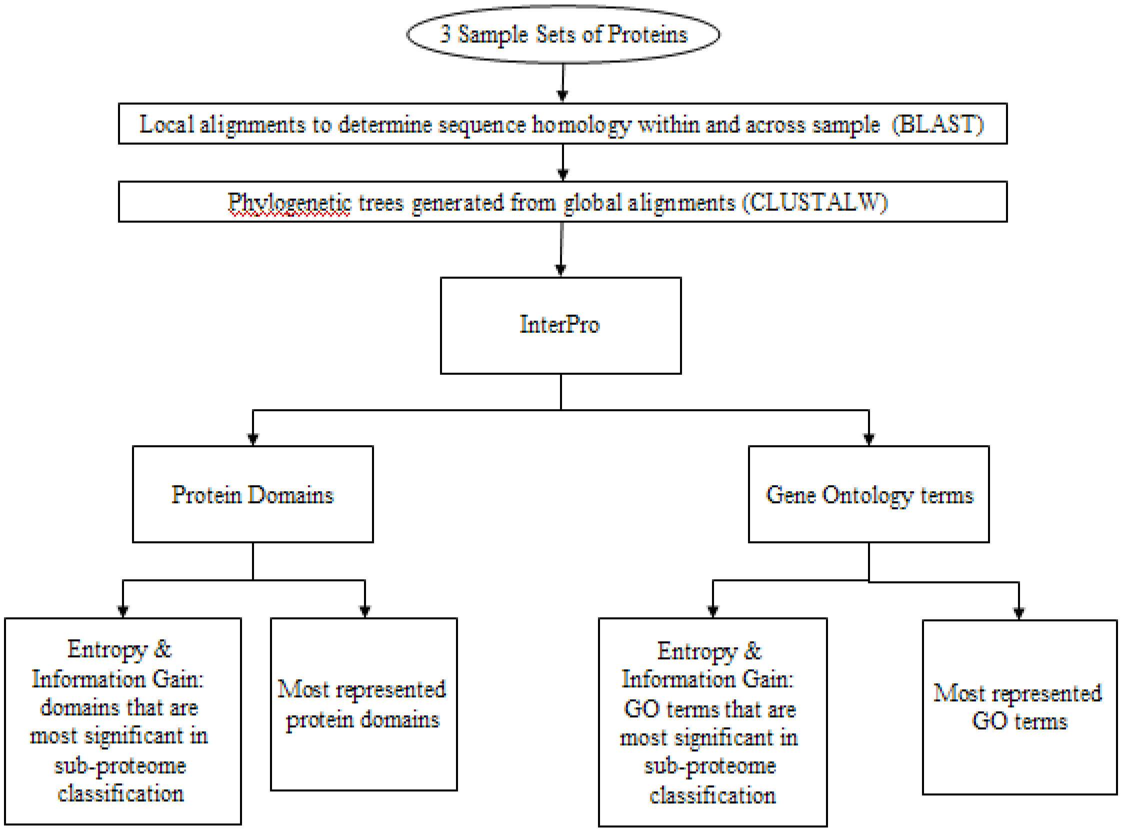

3 Sample Sets of Proteins
Local alignments to determine sequence homology within and across sample (BLAST)
Phylogenetic trees generated from global alignments (CLUSTALW)
InterPro
Protein Domains
Gene Ontology terms
Entropy & Information Gain: domains that are most significant in sub-proteome classification
Most represented protein domains
Entropy & Information Gain: GO terms that are most significant in sub-proteome classification
Most represented GO terms

Figure 8. Workflow #2
Annotation (BLAST, Protein Domain, GO) and statistical analysis of annotation

IV. RESULTS

A. Introduction

Detailed and summary results are provided in the following Specific Findings result section. Results are presented in the following order: 1) signal pepide analysis, 2) results of method validation, 3) a detailed explanation of steps followed in the motif analysis and summary results, 4) homology analysis, 5) protein domain and GO annotation data, 6) statistical analysis of annotation data including the most significant domains and GO terms.

B. Important Highlights

Reiterative use of MEME and MAST yielded a proposed conserved sequence motif of x[K/R]xGE across 11 sample proteins and 2 proteins from *S. japonicum*. The vesicles were enriched for the motif. An effort to reproduce the workflow followed by Hiller *et al,* which resulted in discovery of a conserved motif across *P. falciparum* proteins, was only successful after manual manipulation of the N-terminal starting positions of the initial five protein sequences used in that analysis. The use of a stringent background model was found to be important to the success of MEME searches. Interestingly, the sample proteins had a significantly lower percentage of predicted signal peptides. Methods were developed for quickly creating matrices of protein domains or GO terms *versus* protein names. These matrices provided a way for comparing proteins to determine similarities and differences. They were useful in determining functional differences in the lipid-induced proteins *versus* the tail shearing-induced proteins, and in evaluating proteins proposed to contain a conserved motif. Statistical methods for determining the most statistically significant domains and GO terms were also automated and applied. It was determined that ATP binding was the most represented biological function in the vesicles, at a level 2-1/2 times higher than in the secretions, and glycolysis was the most represented biological process in the secretions.

C. Specific Findings

Signal Peptide Prediction

Table 1 below provides SignalP neural network predictions of signal peptides and cleavage sites for each of the sample sets. Figure 9 gives a summary comparison of the three different sample sets by showing the total number of proteins in each sample set and the total number of proteins that contain a signal peptide in each sample set. All sample sets had relatively low numbers of sequences that were predicted to contain a signal peptide. 11%, 9%, and 7% of vesicles, secretions, and tegument proteins respectively were predicted to have signal peptides. By contrast, 36 of 155 control proteins, or 23% were predicted to contain a signal peptide. It was expected that a higher number of the secreted proteins contained in the vesicles, secretions, and tegument proteins would have a signal peptide than the control proteins. Possible reasons for this discrepancy will be discussed later.

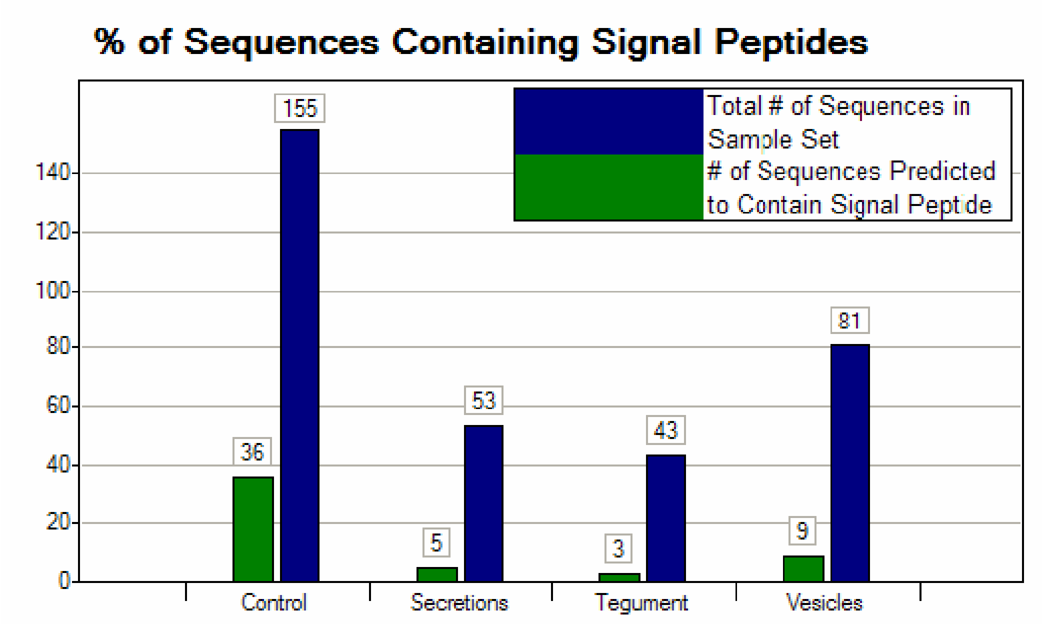


Figure 9. Comparison of 3 Sample Sets

Table 1. Signal peptide predictions

Tables below show SignalP neural network prediction results. Note: Only positive results are shown for the control samples. Two neural networks are used by SignalP: one to predict the actual signal peptide and one to predict the cleavage site. S-score: calculated for every amino acid in the submitted sequence; a high S-score for an amino acid is an indicator that it is part of a signal peptide. S-mean is the average of the S-score and is calculated for the length of the signal peptide. C-max: measure of the likelihood of an amino acid being part of the cleavage site; calculated for every amino acid in the submitted sequence; the amino acids with the highest C-max scores are predicted to be part of the cleavage site. Y-max: derived from a combination of the C-max score and S-mean score. D-score: a new scoring method in version 3.0, the average of the S-mean and Y-max scores. It outperforms S-score for discriminating between signal peptides and non-signal peptides (Bendtsen, Nielsen et al. 2004).

Vesicles

description	**Name**	**Cmax**	**pos**	**?**	**Ymax**	**pos**	**?**	**Smax**	**pos**	**?**	**Smean**	**?**	**D**	**?**
Similar to B-cell receptor-associated protein 32	TC10689	0.103	25	N	0.213	25	N	0.976	12	Y	0.687	Y	0.45	**Y**
Surface protein, fluke	T30271	0.594	20	Y	0.695	20	Y	0.978	2	Y	0.898	Y	0.796	**Y**
Protein-disulfide isomerase homolog	CAA80520	0.284	16	N	0.442	18	Y	0.987	2	Y	0.94	Y	0.691	**Y**
Elastase 2a	AAM43941	0.62	25	Y	0.646	25	Y	0.971	13	Y	0.82	Y	0.733	**Y**
Endoplasmin	AAF66929	0.861	21	Y	0.691	21	Y	0.981	16	Y	0.932	Y	0.811	**Y**
SPO-1 protein (anti-inflammatory protein 6)	AAD26122	0.357	17	Y	0.378	17	Y	0.994	14	Y	0.942	Y	0.66	**Y**
Elastase (elastase 1b)	AAC46967	0.797	25	Y	0.719	25	Y	0.98	6	Y	0.825	Y	0.772	**Y**
Calreticulin	AAA19024	0.86	17	Y	0.831	17	Y	0.958	2	Y	0.879	Y	0.855	**Y**
Pancreatic elastase precursor (elastase 1a)	A28942	0.603	26	Y	0.616	26	Y	0.972	7	Y	0.82	Y	0.718	**Y**
Phosphoglycerate mutase	TC7546	0.072	39	N	0.043	15	N	0.314	2	N	0.085	N	0.064	N
Similar to HEL protein	TC7459	0.109	22	N	0.041	17	N	0.192	1	N	0.051	N	0.046	N
Similar to pyruvate kinase	TC7454	0.31	24	N	0.353	24	Y	0.748	13	N	0.491	Y	0.422	N

Vesicles

description	Name	Cmax	pos	?	Ymax	pos	?	Smax	pos	?	Smean	?	D	?
Weakly similar to troponin T	TC7449	0.174	25	N	0.013	25	N	0.028	2	N	0.013	N	0.013	N
Homolog to tubulin beta-2 chain	TC7336	0.076	19	N	0.047	19	N	0.104	1	N	0.055	N	0.051	N
Similar to T-complex protein-1, epsilon subunit	TC6878	0.075	21	N	0.086	11	N	0.491	1	N	0.136	N	0.111	N
Similar to hypothetical Schistosoma japonicum protein	TC17017	0.095	24	N	0.128	24	N	0.597	3	N	0.274	N	0.201	N
Similar to chaperonin containing T-complex protein-1, zeta subunit	TC16896	0.185	17	N	0.124	30	N	0.238	13	N	0.107	N	0.116	N
Similar to ADP/ATP translocase	TC16858	0.254	25	N	0.3	32	N	0.831	22	N	0.229	N	0.265	N
Similar to malate dehydrogenase precursor	TC16844	0.07	25	N	0.111	22	N	0.696	9	N	0.365	N	0.238	N
Homolog to calmodulin	TC16812	0.1	27	N	0.025	42	N	0.044	7	N	0.016	N	0.02	N
Similar to lactate dehydrogenase	TC16735	0.275	39	N	0.341	39	Y	0.917	35	Y	0.222	N	0.282	N
Similar to transketolase	TC16539	0.123	29	N	0.047	29	N	0.087	1	N	0.038	N	0.042	N
Similar to histone H4	TC14578	0.019	25	N	0.021	4	N	0.096	1	N	0.08	N	0.051	N
Similar to succinate dehydrogenase Fp subunit	TC13974	0.149	23	N	0.042	15	N	0.235	1	N	0.065	N	0.053	N
Similar to chaperonin containing T-complex protein-1, gamma subunit	TC13671	0.277	41	N	0.262	41	N	0.831	30	N	0.122	N	0.192	N
Similar to histone H3	TC13658	0.023	24	N	0.017	11	N	0.073	1	N	0.05	N	0.034	N
Similar to chaperonin containing T-complex protein-1, beta subunit	TC13620	0.066	22	N	0.059	13	N	0.531	1	N	0.149	N	0.104	N
Homolog to H2B histone	TC13606	0.028	48	N	0.017	9	N	0.051	1	N	0.033	N	0.025	N
Similar to ATP synthase [1]-chain mito	TC13604	0.091	27	N	0.095	21	N	0.427	11	N	0.225	N	0.16	N
Similar to glycogen phosphorylase, muscle	TC13591	0.055	32	N	0.029	32	N	0.107	11	N	0.025	N	0.027	N
Similar to nucleoside-diphosphate kinase	TC11413	0.19	20	N	0.023	41	N	0.111	5	N	0.045	N	0.034	N
Thioredoxin peroxidase 3	TC10839	0.057	22	N	0.074	22	N	0.394	12	N	0.173	N	0.124	N
Similar to ribosomal protein L12	TC10765	0.161	20	N	0.085	25	N	0.387	5	N	0.165	N	0.125	N
Similar to citrate synthase	TC10655	0.034	39	N	0.024	39	N	0.08	3	N	0.024	N	0.024	N
Weakly similar to phosphoglucomutase	TC10613	0.259	16	N	0.071	16	N	0.082	1	N	0.041	N	0.056	N

Vesicles

description	Name	Cmax	pos	?	Ymax	pos	?	Smax	pos	?	Smean	?	D	?
Similar to transitional ER ATPase	TC10596	0.044	16	N	0.02	16	N	0.062	11	N	0.027	N	0.024	N
Similar to ATP synthase alpha-chain mito	TC10585	0.044	21	N	0.065	4	N	0.74	1	N	0.612	Y	0.339	N
Weakly similar to heterogeneous nuclear ribonucleoprotein A2 homolog 1	TC10489	0.389	32	Y	0.224	32	N	0.487	29	N	0.078	N	0.151	N
T-complex protein-1, alpha subunit	Q94757	0.087	22	N	0.092	35	N	0.378	31	N	0.118	N	0.105	N
Probable dynein light chain (SM10)	Q94748	0.084	28	N	0.023	28	N	0.061	1	N	0.019	N	0.021	N
Enolase (2-phosphoglycerate dehydratase)	Q27877	0.254	17	N	0.114	17	N	0.232	2	N	0.091	N	0.102	N
Peptidyl-prolyl cis-trans isomerase (PPIase)	Q26565	0.039	32	N	0.043	18	N	0.291	12	N	0.108	N	0.075	N
14-3-3 protein homolog 1	Q26540	0.1	27	N	0.045	27	N	0.113	1	N	0.05	N	0.047	N
Fructose-bisphosphate aldolase	P53442	0.091	23	N	0.055	33	N	0.114	29	N	0.035	N	0.045	N
Triose-phosphate isomerase (TPI)	P48501	0.03	21	N	0.034	5	N	0.196	2	N	0.151	N	0.093	N
Tropomyosin 2 (TMII)	P42638	0.044	23	N	0.022	2	N	0.079	1	N	0.079	N	0.051	N
Tropomyosin 1 (TMI) (polypeptide 49)	P42637	0.036	19	N	0.016	2	N	0.049	1	N	0.049	N	0.032	N
Phosphoglycerate kinase	P41759	0.081	32	N	0.039	32	N	0.124	13	N	0.049	N	0.044	N
GAPDH (major larval surface antigen) (P-37)	P20287	0.078	25	N	0.168	25	N	0.769	11	N	0.508	Y	0.338	N
ATP:guanidino kinase SMC74	P16641	0.039	38	N	0.022	38	N	0.06	1	N	0.017	N	0.02	N
Major egg antigen P40	P12812	0.035	44	N	0.039	5	N	0.225	1	N	0.146	N	0.093	N
Glutathione S-transferase (Sm, 211 aa)	P09792	0.657	18	Y	0.082	18	N	0.17	26	N	0.066	N	0.074	N
Paramyosin	P06198	0.064	23	N	0.027	28	N	0.072	5	N	0.039	N	0.033	N
Similar to myosin regulatory light chain	CD180182	0.309	27	N	0.045	37	N	0.082	36	N	0.032	N	0.038	N
Elongation factor 1-alpha	CAA69721	0.106	20	N	0.026	4	N	0.119	3	N	0.103	N	0.065	N
70000 molecular weight antigen/hsp70 homolog	CAA28976	0.072	33	N	0.029	33	N	0.06	11	N	0.024	N	0.026	N
Actin-binding/filamin-like protein	AAR26703	0.218	44	N	0.065	44	N	0.089	1	N	0.026	N	0.046	N
Arginase	AAP94031	0.295	35	N	0.074	35	N	0.121	1	N	0.029	N	0.051	N
Heat shock protein HSP60	AAM69406	0.242	33	N	0.19	33	N	0.437	29	N	0.097	N	0.144	N
Thioredoxin	AAL79841	0.116	40	N	0.109	40	N	0.33	34	N	0.054	N	0.081	N
Putative histamine-releasing factor	AAL11633	0.366	25	Y	0.035	33	N	0.127	25	N	0.034	N	0.034	N
SNaK1 (Na+/K+-ATPase alpha)	AAL09322	0.027	33	N	0.017	33	N	0.016	3	N	0.011	N	0.014	N

Vesicles

description	Name	Cmax	pos	?	Ymax	pos	?	Smax	pos	?	Smean	?	D	?
14-3-3 epsilon isoform	AAF21436	0.066	27	N	0.032	3	N	0.167	1	N	0.13	N	0.081	N
Myosin light chain	AAD41591	0.297	28	N	0.048	28	N	0.09	9	N	0.035	N	0.042	N
Phosphoenolpyruvate carboxykinase	AAD24794	0.041	30	N	0.024	30	N	0.107	4	N	0.037	N	0.03	N
Thioredoxin peroxidase 2	AAD17299	0.04	55	N	0.025	11	N	0.134	5	N	0.091	N	0.058	N
Tegumental protein Sm20.8	AAC79131	0.212	19	N	0.068	19	N	0.175	1	N	0.056	N	0.062	N
Calcium ATPase 2	AAC72756	0.179	26	N	0.021	26	N	0.072	1	N	0.018	N	0.02	N
Gynecophoral canal protein	AAC47216	0.026	17	N	0.017	37	N	0.065	4	N	0.022	N	0.02	N
Actin 2	AAC46966	0.128	24	N	0.037	38	N	0.057	19	N	0.028	N	0.032	N
Unknown (serpin)	AAB86571	0.127	27	N	0.03	37	N	0.081	4	N	0.024	N	0.027	N
Calponin homolog	AAB47536	0.088	21	N	0.032	3	N	0.177	1	N	0.128	N	0.08	N
Putative cytosol aminopeptidase	AAB41442	0.119	21	N	0.113	21	N	0.648	1	N	0.186	N	0.15	N
glutathione S-transferase, GST [Sm]	AAB21173	0.104	22	N	0.193	22	N	0.879	12	Y	0.372	N	0.282	N
Calcium-binding protein	AAA29921	0.069	19	N	0.017	19	N	0.064	12	N	0.019	N	0.018	N
Fimbrin	AAA29882	0.059	25	N	0.024	36	N	0.057	2	N	0.022	N	0.023	N
Myosin heavy chain	A59287	0.047	24	N	0.02	34	N	0.046	21	N	0.023	N	0.021	N
Tubulin alpha	A48433	0.117	20	N	0.1	20	N	0.345	12	N	0.132	N	0.116	N
Vaccine-dominant antigen Sm21.7	A45630	0.202	26	N	0.045	26	N	0.344	3	N	0.066	N	0.056	N
Heat shock protein 86, fluke (fragment)	A45529	0.032	46	N	0.016	46	N	0.049	3	N	0.015	N	0.016	N
Calcium-binding protein, fluke	A30792	0.078	21	N	0.051	8	N	0.206	1	N	0.065	N	0.058	N

Secretions

description	name	Cmax	pos	?	Ymax	pos	?	Smax	pos	?	Smean	?	D	?
Elastase 1b	AAC46967	0.797	25	Y	0.719	25	Y	0.98	6	Y	0.825	Y	0.772	**Y**
Peptidyl-prolyl cis-trans isomerase B precursor (Sm)	Q26551	0.502	24	Y	0.612	24	Y	0.978	5	Y	0.9	Y	0.756	**Y**
Elastase 2a (Sm)	AAM43941	0.62	25	Y	0.646	25	Y	0.971	13	Y	0.82	Y	0.733	**Y**
Pancreatic elastase precursor (elastase 1a) (Sm)	A28942	0.603	26	Y	0.616	26	Y	0.972	7	Y	0.82	Y	0.718	**Y**
SPO-1 protein (anti-inflammatory protein 6) (Sm)	AAD26122	0.357	17	Y	0.378	17	Y	0.994	14	Y	0.942	Y	0.66	**Y**
Similar to pyruvate kinase (Sm)	TC7454	0.31	24	N	0.353	24	Y	0.748	13	N	0.491	Y	0.422	N
GAPDH (major larval surface antigen) (P-37) (Sm)	P20287	0.078	25	N	0.168	25	N	0.769	11	N	0.508	Y	0.338	N

Secretions														
description	**name**	**Cmax**	**pos**	**?**	**Ymax**	**pos**	**?**	**Smax**	**pos**	**?**	**Smean**	**?**	**D**	**?**
GST (Sm, 218 aa)	AAB21173	0.104	22	N	0.193	22	N	0.879	12	Y	0.372	N	0.282	N
Weakly similar to lactate dehydrogenase (Sm)	TC16735	0.275	39	N	0.341	39	Y	0.917	35	Y	0.222	N	0.282	N
Similar to malate dehydrogenase, mito (Sm)	TC16844	0.07	25	N	0.111	22	N	0.696	9	N	0.365	N	0.238	N
ATP-diphosphohydrolase 1 (Sm)	AAP94734	0.1	24	N	0.139	57	N	0.899	52	Y	0.318	N	0.228	N
Similar to malate dehydrogenase, cytosolic (Sm)	TC17066	0.11	22	N	0.168	22	N	0.685	12	N	0.275	N	0.221	N
Cu,Zn-superoxide dismutase (Sm)	AAC14467	0.074	16	N	0.129	16	N	0.506	4	N	0.263	N	0.196	N
6-Phosphofructokinase (Sm)	Q27778	0.242	30	N	0.08	14	N	0.442	12	N	0.284	N	0.182	N
Similar to ATP synthase _-chain mito (Sm)	TC13604	0.091	27	N	0.095	21	N	0.427	11	N	0.225	N	0.16	N
Heat shock protein HSP60 (Sm)	AAM69406	0.242	33	N	0.19	33	N	0.437	29	N	0.097	N	0.144	N
Putative cytosol aminopeptidase (Sm)	AAB41442	0.127	20	N	0.108	20	N	0.273	10	N	0.14	N	0.124	N
Enolase (2-phosphoglycerate dehydratase) (Sm)	Q27877	0.254	17	N	0.114	17	N	0.232	2	N	0.091	N	0.102	N
Cysteine protease inhibitor (Sm)	AAQ16180	0.051	17	N	0.031	4	N	0.188	3	N	0.16	N	0.096	N
Triose-phosphate isomerase (TIM) (Sm)	P48501	0.03	21	N	0.034	5	N	0.196	2	N	0.151	N	0.093	N
Fatty acid-binding protein Sm14 (Sm)	AAL15461	0.227	22	N	0.066	36	N	0.255	12	N	0.099	N	0.082	N
Thioredoxin (Sm)	AAL79841	0.116	40	N	0.109	40	N	0.33	34	N	0.054	N	0.081	N
Calponin homolog (Sm)	AAB47536	0.088	21	N	0.032	3	N	0.177	1	N	0.128	N	0.08	N
Peptidyl-prolyl cis-trans isomerase (PPIase) (Sm) 1	Q26565	0.039	32	N	0.043	18	N	0.291	12	N	0.108	N	0.075	N
Glutathione *S*-transferase, 28 kDa (GST 28) (Sm)	P09792	0.657	18	Y	0.082	18	N	0.17	26	N	0.066	N	0.074	N
Similar to carbonyl reductase (Sm)	AAC46898	0.16	30	N	0.062	17	N	0.222	12	N	0.078	N	0.07	N
Elongation factor 1 (Sm)	CAA69721	0.106	20	N	0.026	4	N	0.119	3	N	0.103	N	0.065	N
Homolog to phosphoglycerate mutase (Sm)	TC7546	0.072	39	N	0.043	15	N	0.314	2	N	0.085	N	0.064	N
Tegumental protein Sm20.8 (Sm)	AAC79131	0.212	19	N	0.068	19	N	0.175	1	N	0.056	N	0.062	N
Thioredoxin peroxidase 2 (Sm)	AAD17299	0.04	55	N	0.025	11	N	0.134	5	N	0.091	N	0.058	N
Vaccine-dominant antigen Sm21.7 (Sm)	A45630	0.202	26	N	0.045	26	N	0.344	3	N	0.066	N	0.056	N
Ferritin-2 heavy chain (Sm)	P25319	0.156	35	N	0.074	35	N	0.097	33	N	0.033	N	0.053	N
Similar to histone H4 (Sm)	TC14578	0.019	25	N	0.021	4	N	0.096	1	N	0.08	N	0.051	N
Homolog to tubulin _-2 chain (Sm)	TC7336	0.076	19	N	0.047	19	N	0.104	1	N	0.055	N	0.051	N
14-3-3 protein homolog 1 (Sm)	Q26540	0.1	27	N	0.045	27	N	0.113	1	N	0.05	N	0.047	N
Actin-binding/filamin-like protein (Sm)	AAR26703	0.218	44	N	0.065	44	N	0.089	1	N	0.026	N	0.046	N
Fructose-bisphosphate aldolase (Sm)	P53442	0.091	23	N	0.055	33	N	0.114	29	N	0.035	N	0.045	N
Phosphoglycerate kinase (Sm)	P41759	0.081	32	N	0.039	32	N	0.124	13	N	0.049	N	0.044	N
Similar to nucleoside-diphosphate kinase (Sm)	TC11413	0.19	20	N	0.023	41	N	0.111	5	N	0.045	N	0.034	N

Secretions

description	name	Cmax	pos	?	Ymax	pos	?	Smax	pos	?	Smean	?	D	?
Similar to histone H3 (Sm)	TC13658	0.023	24	N	0.017	11	N	0.073	1	N	0.05	N	0.034	N
Actin (Sm)	AAC46966	0.128	24	N	0.037	38	N	0.057	19	N	0.028	N	0.032	N
Phosphoenolpyruvate carboxykinase (Sm)	AAD24794	0.041	30	N	0.024	30	N	0.107	4	N	0.037	N	0.03	N
Unknown (serpin) (Sm)	AAB86571	0.127	27	N	0.03	37	N	0.081	4	N	0.024	N	0.027	N
Similar to muscle glycogen phosphorylase (Sm)	TC13591	0.055	32	N	0.029	32	N	0.107	11	N	0.025	N	0.027	N
70,000 molecular weight antigen/hsp70 homolog (Sm)	CAA28976	0.072	33	N	0.029	33	N	0.06	11	N	0.024	N	0.026	N
Calpain (EC 3.4.22.17) large chain (Sm)	A39343	0.111	17	N	0.025	17	N	0.068	1	N	0.025	N	0.025	N
Homolog to H2B histone (Sm)	TC13606	0.028	48	N	0.017	9	N	0.051	1	N	0.033	N	0.025	N
Fimbrin (Sm)	AAA2988	0.059	25	N	0.024	36	N	0.057	2	N	0.022	N	0.023	N
Probable dynein light chain (SM10) (T-cell-stimulating antigen SM10)	Q94748	0.084	28	N	0.023	28	N	0.061	1	N	0.019	N	0.021	N
ATP:guanidino kinase SMC74 (Sm)	P16641	0.039	38	N	0.022	38	N	0.06	1	N	0.017	N	0.02	N
Homolog to calmodulin (Sm)	TC16812	0.1	27	N	0.025	42	N	0.044	7	N	0.016	N	0.02	N
Calcium-binding protein (Sm)	AAA29921	0.069	19	N	0.017	19	N	0.064	12	N	0.019	N	0.018	N
Heat shock protein 86 (Sm)	A45529	0.032	46	N	0.016	46	N	0.049	3	N	0.015	N	0.016	N

Tegument

description	name	Cmax	pos	?	Ymax	pos	?	Smax	pos	?	Smean	?	D	?
CD63-like protein Sm-TSP-2a	TC13732	0.205	22	N	0.282	36	N	0.951	32	Y	0.815	Y	0.549	**Y**
Nebulette	TC13379	0.328	22	Y	0.413	22	Y	0.88	2	Y	0.674	Y	0.543	**Y**
no significant homology found	TC8556	0.36	22	Y	0.217	43	N	0.997	38	Y	0.747	Y	0.482	**Y**
putative related to F3G5b	TC11571	0.609	20	Y	0.428	20	Y	0.898	12	Y	0.404	N	0.416	N
22K surface membrane antigen	TC10917	0.122	36	N	0.295	36	N	0.95	31	Y	0.527	Y	0.411	N
rat coatamer beta subunit	TC7811	0.195	63	N	0.184	22	N	0.892	11	Y	0.516	Y	0.35	N
no significant homology found	TC19226	0.17	22	N	0.246	22	N	0.615	10	N	0.406	N	0.326	N
syntenin	TC14697	0.799	21	Y	0.371	21	Y	0.512	10	N	0.279	N	0.325	N
actin-binding and severin family group-protein	TC8208	0.237	26	N	0.21	26	N	0.774	25	N	0.336	N	0.273	N
Exocyst complex component Sec10 (hSec10)	TC11347	0.144	29	N	0.206	29	N	0.681	20	N	0.275	N	0.24	N
fatty acid coenzyme A ligase 5	TC17495	0.347	21	Y	0.123	37	N	0.833	32	N	0.344	N	0.234	N
ATP-diphosphohydrolase 1a	TC11432	0.1	24	N	0.139	57	N	0.899	52	Y	0.318	N	0.228	N
SCP-like extracellular proteinb	TC10635	0.689	23	Y	0.228	23	N	0.767	2	N	0.227	N	0.227	N

Tegument

description	name	Cmax	pos	?	Ymax	pos	?	Smax	pos	?	Smean	?	D	?
SGTP4	TC17442	0.216	24	N	0.209	37	N	0.769	32	N	0.234	N	0.222	N
aquaporin-3	TC10637	0.224	33	N	0.218	50	N	0.684	41	N	0.212	N	0.215	N
no significant homology found	TC9780	0.191	16	N	0.166	16	N	0.623	3	N	0.241	N	0.203	N
DEAD (Asp-Glu-Ala-Asp) box polypeptide 1	CD194580	0.067	25	N	0.108	21	N	0.569	7	N	0.291	N	0.2	N
no significant homology found	AI882660	0.214	24	N	0.145	61	N	0.796	50	N	0.244	N	0.195	N
hypothetical proteinb	TC7948	0.634	21	Y	0.186	21	N	0.431	2	N	0.113	N	0.15	N
annexin 11a isoform 2	CD098410	0.429	25	Y	0.164	25	N	0.38	4	N	0.128	N	0.146	N
no significant homology found	TC14238	0.288	32	N	0.153	32	N	0.342	31	N	0.064	N	0.109	N
SCP-like extracellular proteinb	TC10634	0.056	43	N	0.047	5	N	0.331	1	N	0.165	N	0.106	N
ubiquitin/ribosomal fusion proteinb	TC11590	0.257	29	N	0.059	16	N	0.34	2	N	0.153	N	0.106	N
adenylyl cyclase-associated proteinb	TC8265	0.224	26	N	0.122	26	N	0.355	16	N	0.08	N	0.101	N
no significant homology found	TC9174	0.101	39	N	0.105	39	N	0.317	32	N	0.083	N	0.094	N
filamin isoform A	TC17006	0.181	24	N	0.082	24	N	0.204	23	N	0.095	N	0.089	N
similar to Pcmt1-prov protein	TC11206	0.174	25	N	0.066	26	N	0.372	1	N	0.102	N	0.084	N
Alpha-actinin, sarcomeric (F-actin cross linking protein)	TC14047	0.17	25	N	0.073	25	N	0.274	5	N	0.093	N	0.083	N
no significant homology found	TC13203	0.563	22	Y	0.077	22	N	0.229	1	N	0.085	N	0.081	N
amidase	CD157335	0.086	33	N	0.083	33	N	0.282	25	N	0.073	N	0.078	N
dysferlin	CD157335	0.045	57	N	0.041	17	N	0.184	12	N	0.103	N	0.072	N
LIM and SH3 protein 1b	CD182193	0.077	16	N	0.056	16	N	0.196	3	N	0.068	N	0.062	N
ectonucleotide pyrophosphatase/phosphodiesterase 5	TC14339	0.351	23	Y	0.064	23	N	0.173	2	N	0.049	N	0.057	N
no significant homology found	CD192195	0.075	28	N	0.051	36	N	0.113	12	N	0.049	N	0.05	N
thioredoxin peroxidase 1a	TC14049	0.032	55	N	0.025	17	N	0.133	5	N	0.067	N	0.046	N
HLA-B associated transcript 1b	TC7459	0.109	22	N	0.041	17	N	0.192	1	N	0.051	N	0.046	N
no significant homology found	TC18339	0.092	26	N	0.042	26	N	0.188	1	N	0.039	N	0.04	N
Myosin D	TC13851	0.064	44	N	0.042	44	N	0.1	1	N	0.03	N	0.036	N
phosphatidylinositol transfer protein	TC12230	0.124	33	N	0.038	33	N	0.103	2	N	0.028	N	0.033	N
breast adenocarcinoma marker like (26.0 kD) (2N58)	TC13662	0.041	24	N	0.028	24	N	0.088	11	N	0.035	N	0.031	N
major egg antigen	TC7485	0.022	25	N	0.012	6	N	0.059	3	N	0.038	N	0.025	N
fimbrin	TC7585	0.059	25	N	0.024	36	N	0.057	2	N	0.022	N	0.023	N

Tegument

description	name	Cmax	pos	?	Ymax	pos	?	Smax	pos	?	Smean	?	D	?
no significant homology found	BF936329	0.05	26	N	0	1	N	0.967	29	Y	0	N	0	N

Controls

description	name	Cmax	pos	?	Ymax	pos	?	Smax	pos	?	Smean	?	D	?
Female specific 800 protein (FS800)	P16463	0.935	23	Y	0.888	**23**	Y	0.993	13	Y	0.931	Y	0.909	**Y**
13 kDa tegumental antigen Sm13 [Sm]	AAC25419.1	0.979	18	Y	0.88	**18**	Y	0.983	2	Y	0.886	Y	0.883	**Y**
peptidylglycine alpha hydroxylating mono-oxygenase [Sm]	AAO18222.1	0.963	18	Y	0.88	**18**	Y	0.978	4	Y	0.877	Y	0.879	**Y**
insulin receptor protein kinase RTK-2 [Sm]	AAN39120.1	0.977	22	Y	0.865	**22**	Y	0.989	12	Y	0.89	Y	0.878	**Y**
nicotinic acetylcholine receptor alpha subunit precursor [Sm]	AAR84361.1	0.962	20	Y	0.852	**20**	Y	0.964	8	Y	0.888	Y	0.87	**Y**
p48 eggshell protein	AAA29908.1	0.871	20	Y	0.844	**20**	Y	0.985	1	Y	0.893	Y	0.869	**Y**
calreticulin	AAA19024.1	0.86	17	Y	0.831	**17**	Y	0.959	2	Y	0.881	Y	0.856	**Y**
albumin precursor [Sm]	AAL08579.1	0.752	19	Y	0.791	**19**	Y	0.974	1	Y	0.919	Y	0.855	**Y**
carbohydrate-binding calcium-dependent lectin precursor [Sm]	AAX63737.1	0.929	24	Y	0.841	**24**	Y	0.979	6	Y	0.869	Y	0.855	**Y**
LGG	AAB81008.1	0.713	19	Y	0.766	**19**	Y	0.973	4	Y	0.907	Y	0.836	**Y**
unknown [Sm]	AAN17275.1	0.65	19	Y	0.716	**19**	Y	0.989	4	Y	0.923	Y	0.819	**Y**
endoplasmin [Sm]	AAF66929.1	0.861	21	Y	0.687	**21**	Y	0.98	16	Y	0.923	Y	0.805	**Y**
glutathione peroxidase-2 [Sm]	AAU34080.1	0.692	19	Y	0.691	**19**	Y	0.949	3	Y	0.907	Y	0.799	**Y**
acetylcholinesterase [Sm]	AAQ14321.1	0.697	26	Y	0.704	**26**	Y	0.959	11	Y	0.822	Y	0.763	**Y**
developmentally regulated antigen 10.3 precursor [Sm]	AAP13803.1	0.641	25	Y	0.677	**25**	Y	0.962	13	Y	0.81	Y	0.744	**Y**
elastase 2a [Sm]	AAM43941.1	0.62	25	Y	0.641	**25**	Y	0.969	13	Y	0.81	Y	0.726	**Y**
receptor tyrosine kinase [Sm]	AAL67949.1	0.423	24	Y	0.55	**24**	Y	0.984	7	Y	0.901	Y	0.725	**Y**
tyrosinase 1 precursor [Sm]	AAP93838.1	0.575	24	Y	0.623	**24**	Y	0.972	2	Y	0.824	Y	0.723	**Y**
unknown [Sm]	AAN17279.1	0.497	27	Y	0.599	**27**	Y	0.984	8	Y	0.845	Y	0.722	**Y**
CD9-like protein Sm-TSP-1 [Sm]	AAN17278.2	0.323	31	Y	0.502	**31**	Y	0.994	20	Y	0.939	Y	0.721	**Y**
unknown [Sm]	AAB86568.1	0.237	17	N	0.431	**17**	Y	0.969	1	Y	0.921	Y	0.676	**Y**
nicotinic acetylcholine receptor non-alpha subunit precursor [Sm]	AAR84362.1	0.625	24	Y	0.633	**24**	Y	0.901	22	Y	0.719	Y	0.676	**Y**
elastase 2b [Sm]	AAM43942.1	0.274	22	N	0.433	**22**	Y	0.995	13	Y	0.878	Y	0.655	**Y**
ORF-RF2; putative	AAA74696.1	0.635	22	Y	0.283	**32**	N	0.985	25	Y	0.944	Y	0.613	**Y**

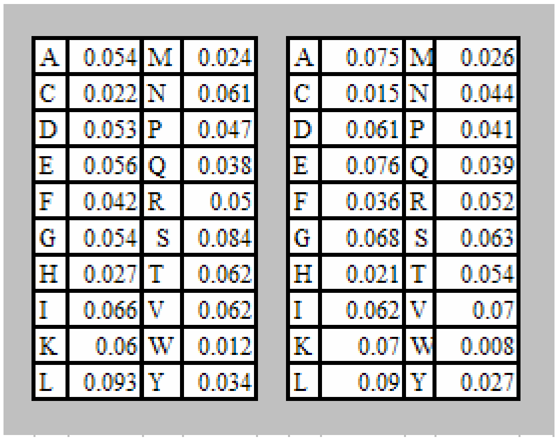
Controls															
description	**name**	**Cmax**	**pos**	**?**	**Ymax**	**pos**	**?**	**Smax**	**pos**	**?**	**Smean**	**?**	**D**	**?**	
cathepsin B endopeptidase [Sm]	CAC85211.2	0.275	28	N	0.414	**23**	Y	0.93	9	Y	0.79	Y	0.602	**Y**	
ORF 2	AAA29910.1	0.267	29	N	0.321	**21**	N	0.945	5	Y	0.812	Y	0.566	**Y**	
CD63-like protein Sm-TSP-2 [Sm]	AAN17276.1	0.205	22	N	0.282	**36**	N	0.951	32	Y	0.816	Y	0.549	**Y**	
NADH dehydrogenase subunit 4L [Sm]	AAG13165.2	0.226	33	N	0.271	**33**	N	0.989	6	Y	0.817	Y	0.544	**Y**	
NADH dehydrogenase subunit 4L [Sm]	NP_66213.2	0.226	33	N	0.271	**33**	N	0.989	6	Y	0.817	Y	0.544	**Y**	
cathepsin B1 isotype 2 [Sm]	CAD44625.1	0.474	18	Y	0.485	**18**	Y	0.867	4	N	0.582	Y	0.534	**Y**	
unknown	AAC46888.1	0.404	30	Y	0.41	**30**	Y	0.964	12	Y	0.639	Y	0.524	**Y**	
Sm29 [Sm]	AAC98911.1	0.665	27	Y	0.478	**27**	Y	0.856	6	N	0.539	Y	0.509	**Y**	
neuropeptide F precursor [Sm]	AAT77204.1	0.279	39	N	0.489	**39**	Y	0.993	28	Y	0.501	Y	0.495	**Y**	
potassium channel protein	AAC37227.1	0.785	20	Y	0.51	**20**	Y	0.714	1	N	0.415	N	0.463	**Y**	
alpha 38720 fucosyltransferase [Sm]	AAF71198.1	0.209	35	N	0.394	**35**	Y	0.984	30	Y	0.53	Y	0.462	**Y**	
putative seven transmembrane receptor [Sm]	AAR84066.2	0.089	24	N	0.208	**31**	N	0.932	25	Y	0.688	Y	0.448	**Y**	

Method Validation

To validate the methodology used by Hiller *et al.* (Hiller, Bhattacharjee et al. 2004) to find and refine a conserved sequence motif, an attempt was made to reproduce the authors' workflow using the same exact samples and methods. A pattern had initially been identified by using MEME to analyze the first 100 residues of five *P. falciparum* sequences, known to be vacuolar transport sequences, starting immediately after the predicted SS cleavage site. To validate this, the five sequences were downloaded from the nonredundant protein database and submitted to SignalP for signal peptide prediction. Interestingly, only three of the five sequences were found to contain a signal sequence using the neural networks method (and only two of five using the hidden Markov model method). When just these three sequences were trimmed to 100 residues after the predicted cleavage site and submitted to MEME, the same motif found by Hiller *et al.* was not found. The other two sequences were added into the analysis and submitted to MEME in an attempt to discover the same motif. But only after manipulation of the starting point of the sequences was the same motif found. The rest of the workflow was not difficult to validate. This leads to two points: 1) finding conserved motifs with MEME when the initial number of input sequences is small, is not trivial, and 2) finding a motif using MEME is likely to be more successful when it is complimented by lab experiments. Hiller *et al.* had the benefit of prior experimental knowledge of exactly which amino acids were required for a protein to be exported.

Conserved Motif Analysis

By default, MEME uses the input sequences to build a 0-order Markov model of amino acid frequencies for use as a background model in determining the likelihood of particular motifs. It follows that when the number of input sequences is small, the background model is not as sensitive as when the number of input sequences is large. To improve the sensitivity of the search and provide a consistent background for each analysis, a background model was built based on amino acid frequencies of the control set. The control set consisted of 155 full-length *S. mansoni* cDNA sequences downloaded from NCBI, from which redundancy and highly homologous sequences had been removed. This model was specified as the background in all MEME searches in place of the input sequences. In the control set, shown on the left in Figure 10, amino acids Cys (C), Phe (F), His (H), Ile (I), Asn (N), Ser (S), Thr (T), Trp (W), and Tyr (Y) are enriched, with Asn (1.7% difference) and Ser (2.1% difference) being the most significantly different. In the vesicle data set, shown on the right in Figure 10 below, amino acids Ala (A), Asp (D), Glu (E), Gly (G), Lys (K), and Val (V), with Ala (2.1% difference) and Glu (2.0% difference) being the most significantly different.



A	0.054	M	0.024
C	0.022	N	0.061
D	0.053	P	0.047
E	0.056	Q	0.038
F	0.042	R	0.05
G	0.054	S	0.084
H	0.027	T	0.062
I	0.066	V	0.062
K	0.06	W	0.012
L	0.093	Y	0.034

A	0.075	M	0.026
C	0.015	N	0.044
D	0.061	P	0.041
E	0.076	Q	0.039
F	0.036	R	0.052
G	0.068	S	0.063
H	0.021	T	0.054
I	0.062	V	0.07
K	0.07	W	0.008
L	0.09	Y	0.027

Figure 10. *S. mansoni* amino acid frequencies
Left: Amino acid frequencies of control data set,
used as background in MEME analysis;
Right: Amino acid frequencies of vesicles data set

Without either elimination or careful weighting of overrepresented proteins, the motif analysis led to false positives that were due to high sequence homology, not necessarily to common secretion function. Sequences that had high homology to other input sequences were weighted equally by specifying weights in the first line of the fasta input file as follows. For example, if two highly homologous sequences existed in the file, the first line of the input file contained the line: “>WEIGHTS 0.5 0.5” followed by the two highly homologous then all other sequences.

In order to increase the sensitivity of the search when the number of input sequences to MEME was small (twenty or fewer), parameters were set to assume that each sequence in the data set contained *exactly one* occurrence of each motif and to search for a motif that is five amino acids wide, corresponding to the following example command line: meme –inputfile.fasta –bfile Sm_nr_155.txt –mod oops –w 5 >output.html. The increased speed and sensitivity of this search sacrifices accuracy and may result in fuzzy motifs. When the number of input sequences was not small (greater than twenty), MEME parameters were set to assume that each sequence contains *at most one* occurrence of each motif (since it is likely that some may not contain the motif) and to allow MEME to automatically choose the best motif width (the default). These settings correspond to the example command line: <meme –inputfile.fasta –bfile Sm_nr_155.txt –mod zoops >output.html>. While this option is slower and slightly less sensitive to weak motifs, it results in more accurate motifs.

MEME output includes (as labeled in Figure 11): A) an information content (IC) diagram, B) a simplified PSPM, C) a multilevel consensus sequence, D) a block diagram of the occurrence of the motif in each training/input sequence, and E) the occurrences of the motif. In the IC diagram, the positions with the highest IC are the most conserved. The most hydrophobic amino acids are colored blue, while polar, non-charged, non-aliphatic residues are colored green, acidic residues are magenta, and positively charged are red. Other residues are various other colors. The simplified PSPM gives the probability of each possible letter appearing at each possible position in an occurrence of the motif. For readability the letter probabilities are multiplied by 10 and rounded to the nearest integer, where 10 is replaced by “a” and zeros are replaced by ":". The multilevel consensus sequence is calculated from the PSPM. The most likely motif is on the first line and any letter substitutions that have a probability greater than .2 are shown in lines below it. The block diagram shows all occurrences of the motif in each sequence, sorted by the lowest p-value. It can be used to compare relative motif positions across sequences. The p-value of an occurrence is the probability of a single random subsequence the length of the motif, generated according to the 0-order background model, having a score at least as high as the score of the occurrence. The diagram of the occurrences of the motif within each sequence shows the exact motif found in each sequence. It can be used to get a view of how conserved the motif is across sequences.

When the nine vesicle proteins that were predicted to contain an ER-type signal sequence were submitted to MEME, a conserved motif of [E/L]K[H/R]GE resulted, with the K, G, and E residues being the most conserved, as shown in Figure 11, MEME #1. The input sequences (Table 2) included 1) three highly homologous elastases, 2) an endoplasmin and a surface protein that showed 20% homology in a BLAST comparison, 3) four other proteins that had no homology to any other sequences within this subset or the entire vesicle sample set: calreticulin, PDI, a SPO-1 protein, and an unidentified protein. They were trimmed to start at their predicted signal peptide cleavage site and extend 100 amino acids. The three elastases which shared 80% to 92% sequence identity were each weighted equally at 33.3% to eliminate bias caused by overrepresentation of these proteins.

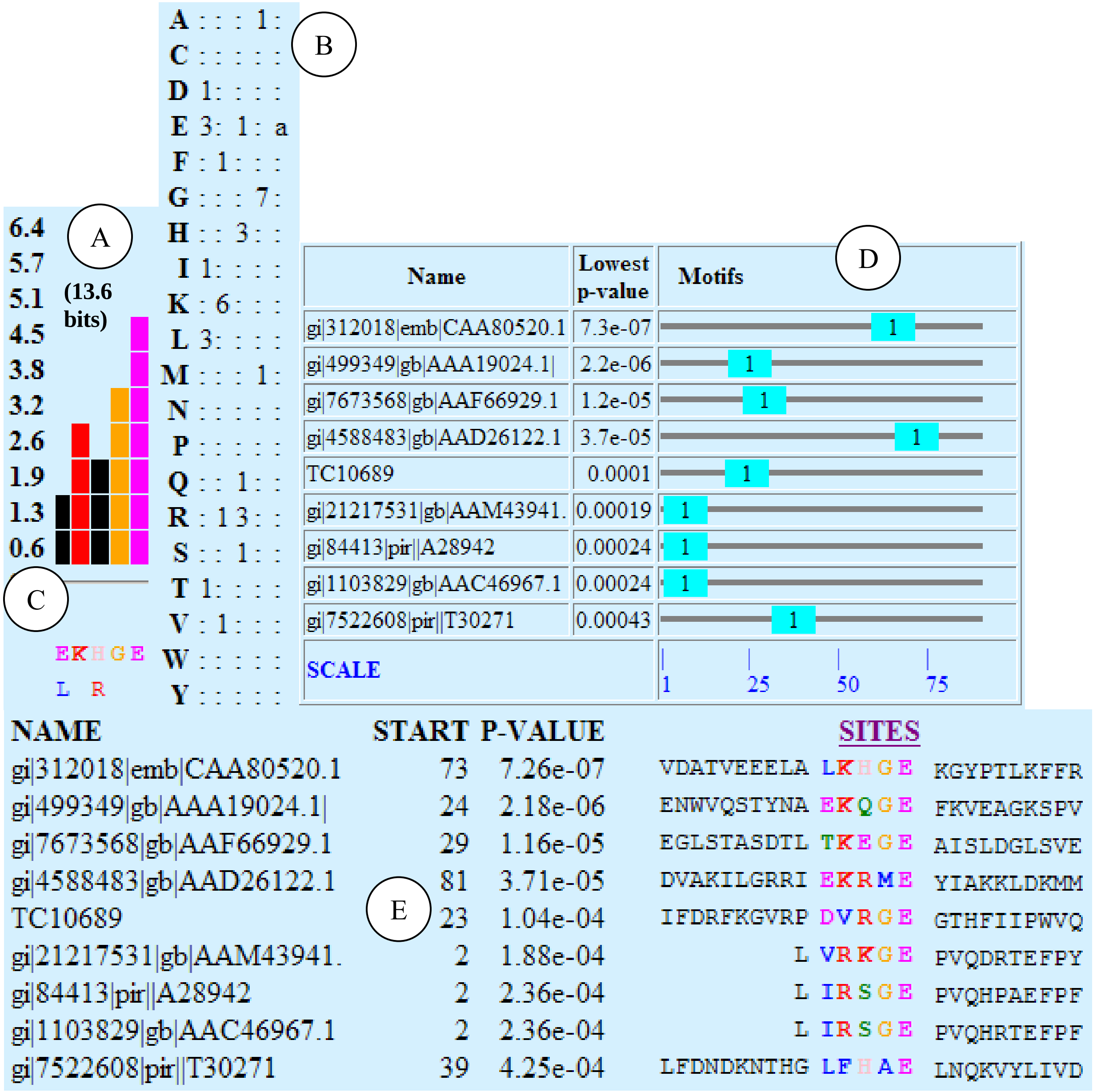
The motif elucidated in MEME #1 contains 13.6 bits of information, an e-value of 770 and a log likelihood value of 85. The motif starts with a positively charged residue, lysine (K), and ends with an acidic residue, glutamic acid (E). Inspection of the occurrences of the motifs shows that it exists across calreticulin, endoplasmin, PDI, and the elastases. Arginine substitutes for lysine in the elastases. The motif does not exist in the SPO-1 protein (AAD26122), the surface protein (T30271), or the protein annotated as similar to B-cell receptor associated protein (T10689).

Table 2. MEME #1 input sequences

accession	Description	weight	source	Signal Sequence?	Trimmed to start at SS and extend 100 aa?	In Sample Sets?
AAC46967	Elastase	0.33	vesicles with signal sequence	no	yes	yes
AAM43941	elastase 2a [Sm]	0.33	vesicles with signal sequence	yes	yes	yes
A28942	pancreatic elastase precursor - fluke (Sm)	0.33	vesicles with signal sequence	yes	yes	yes
AAA19024	calreticulin	1	vesicles with signal sequence	yes	yes	yes
AAF66929	endoplasmin [Sm]	1	vesicles with signal sequence	yes	yes	yes
T30271	surface protein - fluke (Sm)	1	vesicles with signal sequence	yes	yes	yes
CAA80520	PDI homologue [Sm]	1	vesicles with signal sequence	yes	yes	yes
AAD26122	SPO-1 protein [Sm]	1	vesicles with signal sequence	yes	yes	yes
TC10689	Similar to B-cell receptor-associated protein 32	1	vesicles with signal sequence	yes	yes	yes

Figure 11. MEME #1

xKxGE motif: information content = 13.6; e-value = 770; log likelihood ratio = 85



A : : : 1 :
C : : : : :
D 1 : : : :
E 3 : 1 : a
F : 1 : : :
G : : : 7 :
H : : 3 : :
I 1 : : : :
K : 6 : : :
L 3 : : : :
M : : : 1 :
N : : : : :
P : : : : :
Q : : 1 : :
R : 1 3 : :
S : : 1 : :
T 1 : : : :
V : 1 : : :
W : : : : :
Y : : : : :

Name	Lowest p-value	Motifs
gi|312018|emb|CAA80520.1	7.3e-07	1
gi|499349|gb|AAA19024.1|	2.2e-06	1
gi|7673568|gb|AAF66929.1	1.2e-05	1
gi|4588483|gb|AAD26122.1	3.7e-05	1
TC10689	0.0001	1
gi|21217531|gb|AAM43941.	0.00019	1
gi|84413|pir||A28942	0.00024	1
gi|1103829|gb|AAC46967.1	0.00024	1
gi|7522608|pir||T30271	0.00043	1
SCALE		1 25 50 75

NAME	START	P-VALUE	SITES		
gi|312018|emb|CAA80520.1	73	7.26e-07	VDATVEEELA	LKHGE	KGYPTLKFFR
gi|499349|gb|AAA19024.1|	24	2.18e-06	ENWVQSTYNA	EKQGE	FKVEAGKSPV
gi|7673568|gb|AAF66929.1	29	1.16e-05	EGLSTASDTL	TKEGE	AISLDGLSVE
gi|4588483|gb|AAD26122.1	81	3.71e-05	DVAKILGRRI	EKRME	YIAKKLDKMM
TC10689	23	1.04e-04	IFDRFKGVRP	DVRGE	GTHFIIPWVQ
gi|21217531|gb|AAM43941.	2	1.88e-04	L	VRKGE	PVQDRTEFPY
gi|84413|pir||A28942	2	2.36e-04	L	IRSGE	PVQHPAEFPF
gi|1103829|gb|AAC46967.1	2	2.36e-04	L	IRSGE	PVQHRTEFPF
gi|7522608|pir||T30271	39	4.25e-04	LFDNDKNTHG	LFHAE	LNQKVYLIVD

The first MAST search using this motif was against a database of 455 full-length *S. mansoni* proteins downloaded from NCBI's nonredundant protein database. A database containing 455 sequences requires that the motif has at least $\log_2(455)$ bits of information (as discussed on page 42). This corresponds to 8.8 bits, hence the motif is significant enough for the search. Results of MAST search #1 are shown in Figure 12.

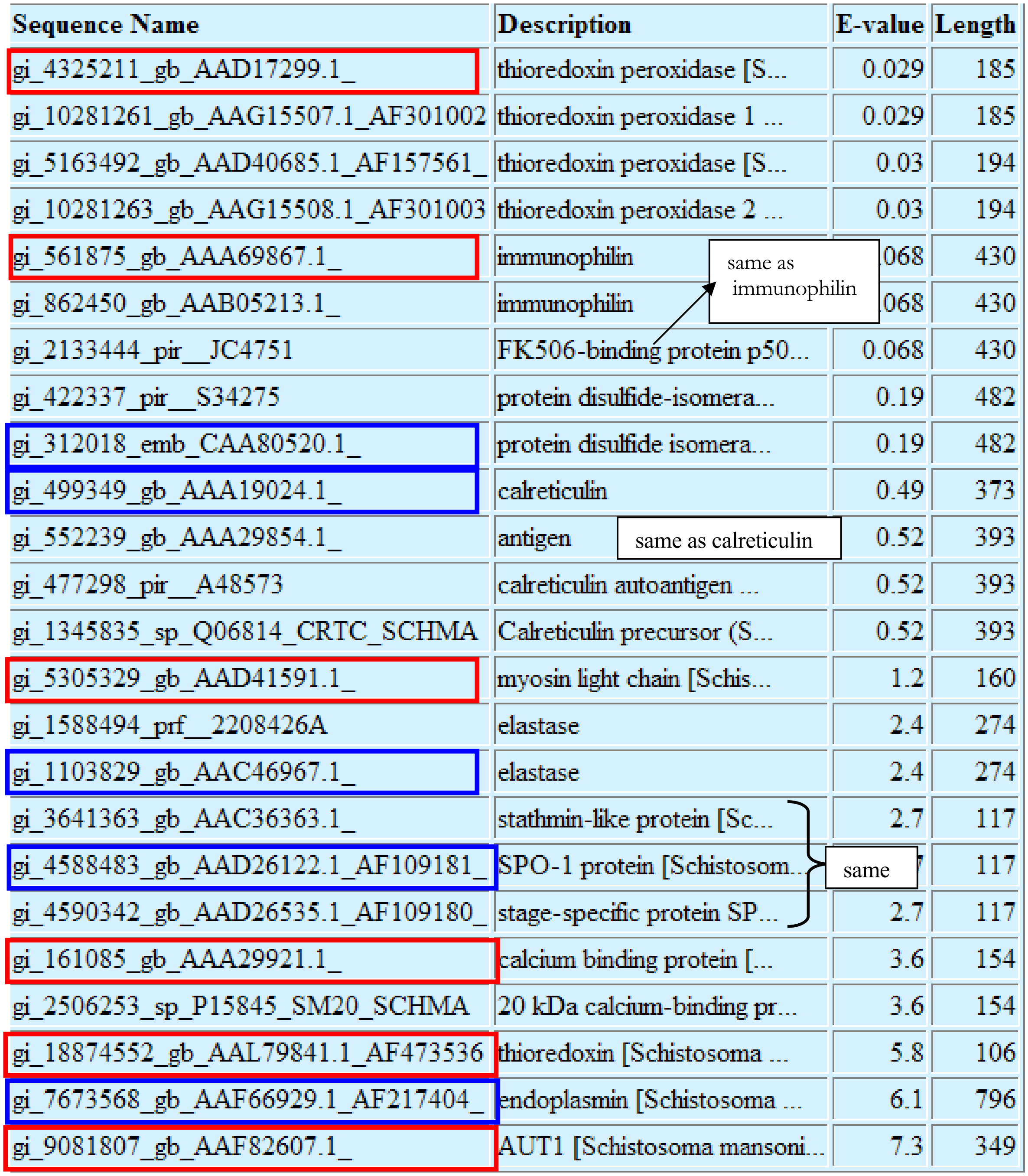
MAST motif diagrams show all motif occurrences. A motif occurrence is defined as a position in the sequence whose match to the motif has position p-value less than .0001. The *position* p-value is equal to the probability that a single random *subsequence of the length of the motif* scoring at least as well as the observed match. The score is computed by summing the appropriate entry from each column of the PSSM. The *sequence* p-value of a score is defined as the probability of a *random sequence of the same length* containing some match with as good or better score. The combined p-value of a sequence (p-value of the product of the sequence p-values) measures the strength of the match of the sequence to all motifs. The e-value of a sequence (combined p-value of the sequence times the number of sequences in the database) is the expected number of sequences in a random database of the same size that would match the motifs as well as the sequence does and is equal to the.

Figure 12 shows that there were 24 sequences with e-values less than 10 resulting from the first MAST search (MAST #1). This included eleven different proteins. Outlined in blue are five proteins that existed in the original data set used in MEME #1 and outlined in red are six new proteins that did not. The new ones include: thioredoxin peroxidase, thioredoxin, myosin light chain, calcium binding protein, immunophilin, and AUT1. Thioredoxin peroxidase is found across all sample sets, thioredoxin is unique to the secretions, myosin light chain is unique to vesicles, calcium binding protein exists in both vesicles and secretions, and immunophilin and AUT1 are not contained in any of the sample sets. Figure 13 depicts relative positions of motifs in order of increasing e-value. The motif occurs in a different position for each protein.

Figure 12. MAST #1 high scoring sequences

Input: motif from MEME #1 & 455 full-length *S. mansoni* proteins NCBI's nonredundant protein database

Proteins not contained in original input sequences (used in MEME #1) are outlined in red, proteins contained in the original input sequences are outlined in blue. Redundant proteins are not outlined.



Sequence Name	**Description**	**E-value**	**Length**
gi_4325211_gb_AAD17299.1_	thioredoxin peroxidase [S...	0.029	185
gi_10281261_gb_AAG15507.1_AF301002	thioredoxin peroxidase 1 ...	0.029	185
gi_5163492_gb_AAD40685.1_AF157561_	thioredoxin peroxidase [S...	0.03	194
gi_10281263_gb_AAG15508.1_AF301003	thioredoxin peroxidase 2 ...	0.03	194
gi_561875_gb_AAA69867.1_	immunophilin	068	430
gi_862450_gb_AAB05213.1_	immunophilin	068	430
gi_2133444_pir__JC4751	FK506-binding protein p50...	0.068	430
gi_422337_pir__S34275	protein disulfide-isomera...	0.19	482
gi_312018_emb_CAA80520.1_	protein disulfide isomera...	0.19	482
gi_499349_gb_AAA19024.1_	calreticulin	0.49	373
gi_552239_gb_AAA29854.1_	antigen	0.52	393
gi_477298_pir__A48573	calreticulin autoantigen ...	0.52	393
gi_1345835_sp_Q06814_CRTC_SCHMA	Calreticulin precursor (S...	0.52	393
gi_5305329_gb_AAD41591.1_	myosin light chain [Schis...	1.2	160
gi_1588494_prf__2208426A	elastase	2.4	274
gi_1103829_gb_AAC46967.1_	elastase	2.4	274
gi_3641363_gb_AAC36363.1_	stathmin-like protein [Sc...	2.7	117
gi_4588483_gb_AAD26122.1_AF109181_	SPO-1 protein [Schistosom...	7	117
gi_4590342_gb_AAD26535.1_AF109180_	stage-specific protein SP...	2.7	117
gi_161085_gb_AAA29921.1_	calcium binding protein [...	3.6	154
gi_2506253_sp_P15845_SM20_SCHMA	20 kDa calcium-binding pr...	3.6	154
gi_18874552_gb_AAL79841.1_AF473536	thioredoxin [Schistosoma ...	5.8	106
gi_7673568_gb_AAF66929.1_AF217404_	endoplasmin [Schistosoma ...	6.1	796
gi_9081807_gb_AAF82607.1_	AUT1 [Schistosoma mansoni...	7.3	349

same as immunophilin

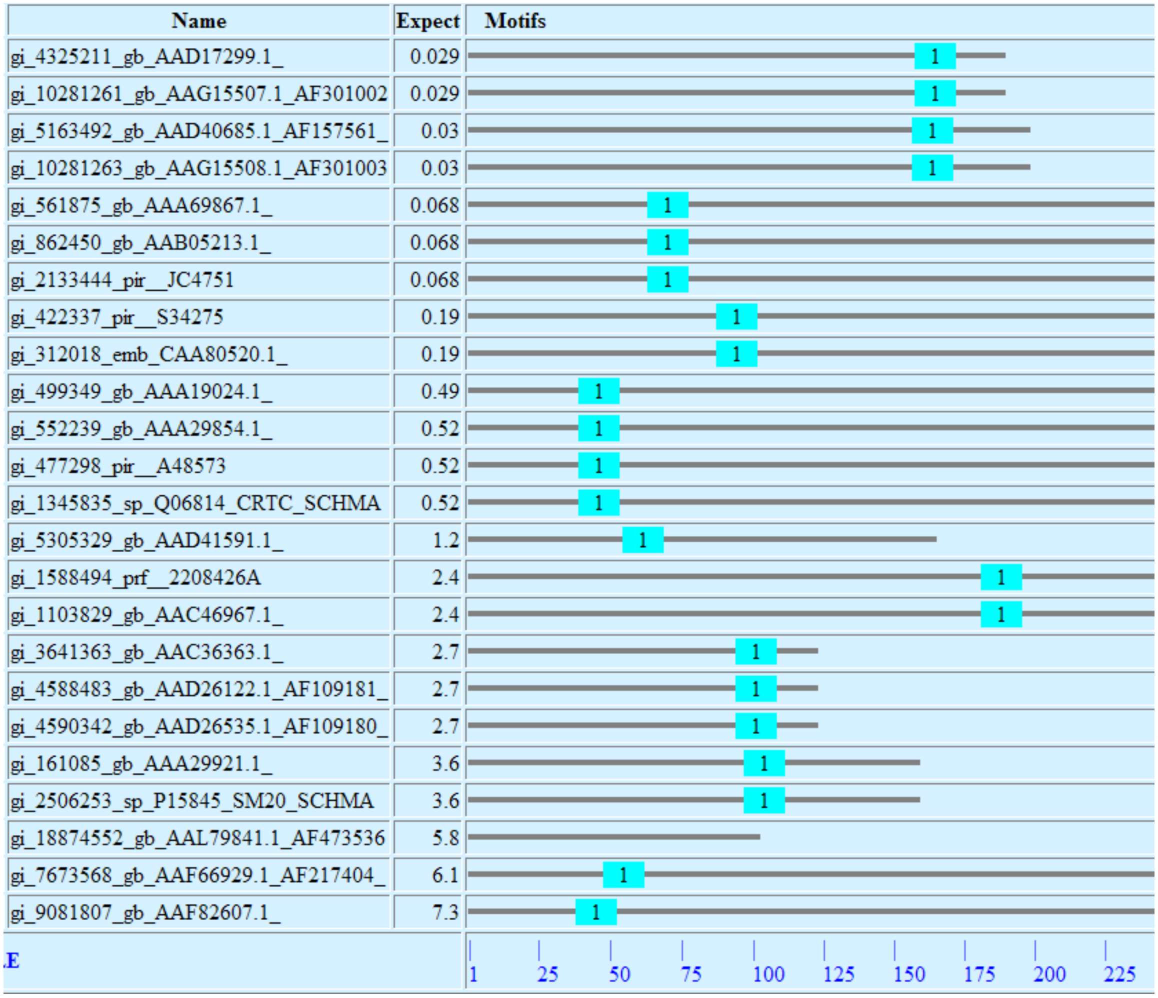
same as calreticulin

same

Figure 13. MAST #1 motif diagram

Input: motif from MEME #1 & 455 full-length *S. mansoni* proteins NCBI's nonredundant protein database

Motifs with significant position p-values (<.0001) are indicated by "1".



Name	**Expect**	**Motifs**
gi_4325211_gb_AAD17299.1_	0.029	1
gi_10281261_gb_AAG15507.1_AF301002	0.029	1
gi_5163492_gb_AAD40685.1_AF157561_	0.03	1
gi_10281263_gb_AAG15508.1_AF301003	0.03	1
gi_561875_gb_AAA69867.1_	0.068	1
gi_862450_gb_AAB05213.1_	0.068	1
gi_2133444_pir__JC4751	0.068	1
gi_422337_pir__S34275	0.19	1
gi_312018_emb_CAA80520.1_	0.19	1
gi_499349_gb_AAA19024.1_	0.49	1
gi_552239_gb_AAA29854.1_	0.52	1
gi_477298_pir__A48573	0.52	1
gi_1345835_sp_Q06814_CRTC_SCHMA	0.52	1
gi_5305329_gb_AAD41591.1_	1.2	1
gi_1588494_prf__2208426A	2.4	1
gi_1103829_gb_AAC46967.1_	2.4	1
gi_3641363_gb_AAC36363.1_	2.7	1
gi_4588483_gb_AAD26122.1_AF109181_	2.7	1
gi_4590342_gb_AAD26535.1_AF109180_	2.7	1
gi_161085_gb_AAA29921.1_	3.6	1
gi_2506253_sp_P15845_SM20_SCHMA	3.6	1
gi_18874552_gb_AAL79841.1_AF473536	5.8	
gi_7673568_gb_AAF66929.1_AF217404_	6.1	1
gi_9081807_gb_AAF82607.1_	7.3	1

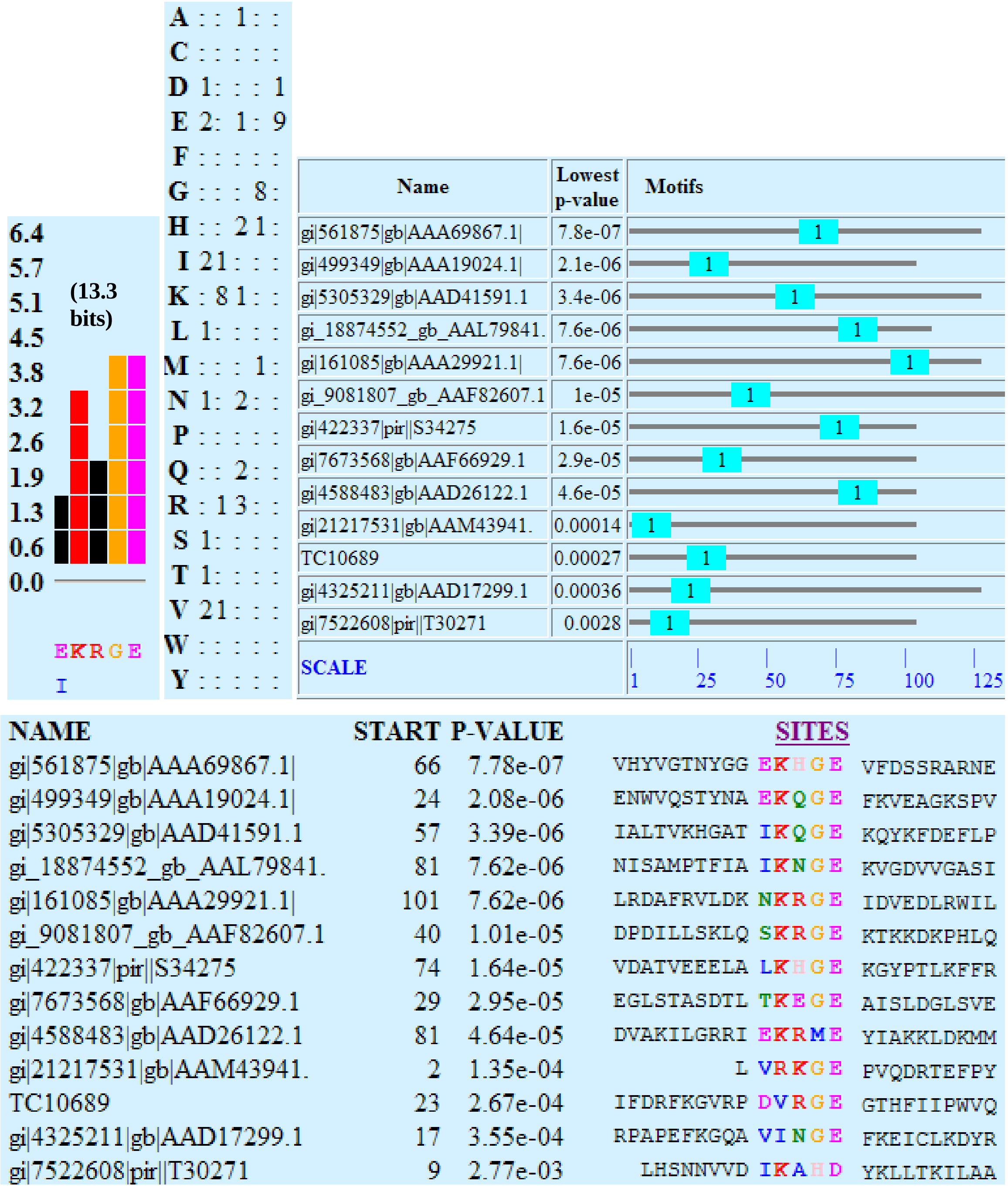
To refine the proposed motif using an additional iteration of MEME, the six new proteins (outlined in red in Figure 12) were added to the set analyzed in MEME #1 (vesicle proteins containing signal peptides). Two of the three redundant elastases were removed from the input sequences so that the input contained only unique sequences. If sequences contained a signal peptide, they were trimmed to start at the signal peptide and extend 100 amino acids. If they did not contain a signal peptide, the entire sequence was used. Sequences used as input to MEME #2 are shown in Table 3 below.

Table 3. MEME #2 input sequences

accession	**description**	**weight**	**source**	**Signal Sequence?**	**Trimmed to start at SS and extend 100 aa?**	**In Sample Sets?**
AAM43941	elastase 2a [Sm]	1	vesicles with signal sequence	yes	yes	yes
AAA19024	calreticulin	1	vesicles with signal sequence	yes	yes	yes
AAF66929	endoplasmin [Sm]	1	vesicles with signal sequence	yes	yes	yes
T30271	surface protein - fluke (Sm)	1	vesicles with signal sequence	yes	yes	yes
CAA80520	PDI homologue [Sm]	1	vesicles with signal sequence	yes	yes	yes
AAD26122	SPO-1 protein [Sm]	1	vesicles with signal sequence	yes	yes	yes
TC10689	Similar to B-cell receptor-associated protein 32	1	vesicles with signal sequence	yes	yes	yes
AAD17299	thioredoxin peroxidase	1	MAST #1- 455 nr Sm sequences	no	no	yes
AAA69867	immunophilin	1	MAST #1- 455 nr Sm sequences	no	no	no
AAD41591	myosin light chain	1	MAST #1- 455 nr Sm sequences	no	no	yes
AAA29921	calcium binding protein	1	MAST #1- 455 nr Sm sequences	no	no	yes
AAL79841	thioredoxin	1	MAST #1- 455 nr Sm sequences	no	no	yes
gi_9081807	AUT1 [Sm]	1	MAST #1- 455 nr Sm sequences	no	no	no

Figure 14. MEME #2

xKxGE motif: information content = 13.3 bits; e-value = .0059; log likelihood ratio = 120



```
A : : 1 : :
C : : : : :
D 1 : : : 1
E 2 : 1 : 9
F : : : : :
G : : : 8 :
H : : 2 1 :
I 2 1 : : :
K : 8 1 : :
L 1 : : : :
M : : : 1 :
N 1 : 2 : :
P : : : : :
Q : : 2 : :
R : 1 3 : :
S 1 : : : :
T 1 : : : :
V 2 1 : : :
W : : : : :
Y : : : : :
```

Name	Lowest p-value	Motifs
gi|561875|gb|AAA69867.1|	7.8e-07	1
gi|499349|gb|AAA19024.1|	2.1e-06	1
gi|5305329|gb|AAD41591.1	3.4e-06	1
gi_18874552_gb_AAL79841.	7.6e-06	1
gi|161085|gb|AAA29921.1|	7.6e-06	1
gi_9081807_gb_AAF82607.1	1e-05	1
gi|422337|pir||S34275	1.6e-05	1
gi|7673568|gb|AAF66929.1	2.9e-05	1
gi|4588483|gb|AAD26122.1	4.6e-05	1
gi|21217531|gb|AAM43941.	0.00014	1
TC10689	0.00027	1
gi|4325211|gb|AAD17299.1	0.00036	1
gi|7522608|pir||T30271	0.0028	1
SCALE		1 25 50 75 100 125

NAME	START	P-VALUE	SITES
gi|561875|gb|AAA69867.1|	66	7.78e-07	VHYVGTNYGG EKHGE VFDSSRARNE
gi|499349|gb|AAA19024.1|	24	2.08e-06	ENWVQSTYNA EKQGE FKVEAGKSPV
gi|5305329|gb|AAD41591.1	57	3.39e-06	IALTVKHGAT IKQGE KQYKFDEFLP
gi_18874552_gb_AAL79841.	81	7.62e-06	NISAMPTFIA IKNGE KVGDVVGASI
gi|161085|gb|AAA29921.1|	101	7.62e-06	LRDAFRVLDK NKRGE IDVEDLRWIL
gi_9081807_gb_AAF82607.1	40	1.01e-05	DPDILLSKLQ SKRGE KTKKDKPHLQ
gi|422337|pir||S34275	74	1.64e-05	VDATVEEELA LKHGE KGYPTLKFFR
gi|7673568|gb|AAF66929.1	29	2.95e-05	EGLSTASDTL TKEGE AISLDGLSVE
gi|4588483|gb|AAD26122.1	81	4.64e-05	DVAKILGRRI EKRME YIAKKLDKMM
gi|21217531|gb|AAM43941.	2	1.35e-04	L VRKGE PVQDRTEFPY
TC10689	23	2.67e-04	IFDRFKGVRP DVRGE GTHFIIPWVQ
gi|4325211|gb|AAD17299.1	17	3.55e-04	RPAPEFKGQA VINGE FKEICLKDYR
gi|7522608|pir||T30271	9	2.77e-03	LHSNNVVD IKAHD YKLLTKILAA

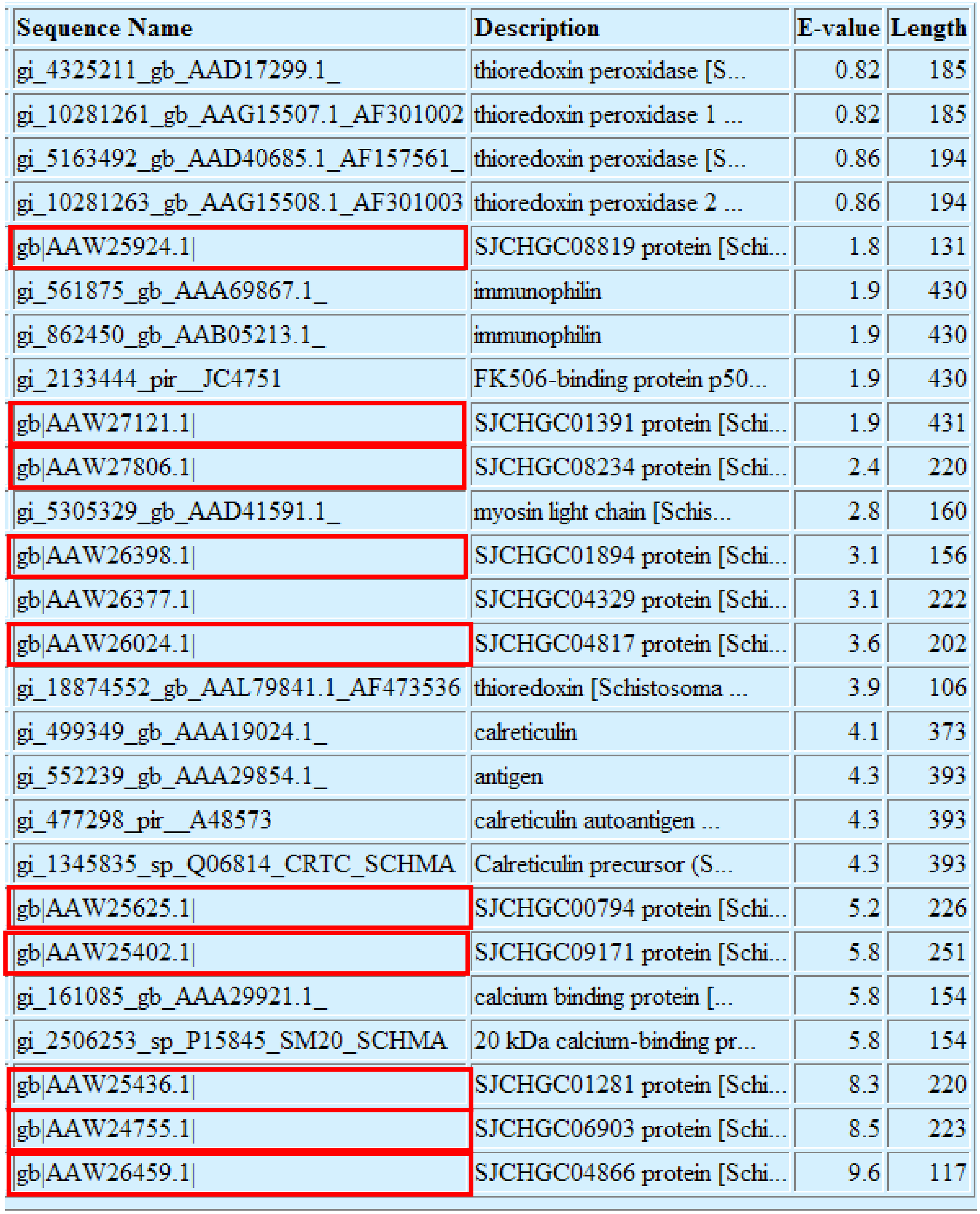
Results of MEME #2 are shown in Figure 14. The consensus sequence changed slightly but the most conserved residues remained unchanged: xKxGE. This motif contains 13.3 bits of information and had an e-value of .0059 (improved from 770 in MEME #1) and a log likelihood ratio of 120 (improved from 85 in MEME #1). Given that for a motif to be considered real it needs to have an e-value below .01, this motif can be considered valid. 13.3 bits of information is sufficient to search databases containing approximately 10,000 or fewer sequences ($\log_2(10000)=13.3$ bits), so this motif could not be used to search the entire SMGI or same, or nr databases, but it could be used to search smaller sets of proteins.

Two additional MAST searches were performed. MAST #2 (Figures 15 and 16) was executed using the motif from MEME #2 to search a subset of NCBI's nonredundant protein database containing ~4000 full-length *S. mansoni* and *S. japonicum* sequences including the 455 *S. mansoni* sequences contained in the database used in the first MAST search. The motif was found in many of the same proteins as in MAST #1: thioredoxin, thioredoxin peroxidase, myosin light chain, calreticulin, calcium binding protein, and immunophilin. It was also found in 10 *S. japonicum* sequences. AAW26398 has 97% identity to myosin light chain which is found in vesicles (AAD41591), AAW25436 has 88% identity to *S. mansoni* thioredoxin peroxidase 3 (AAG15509), AAW25625 is *S. japonicum* thioredoxin peroxidase 2 (BAD90102), and AAW27121 has 78% identity to *S. mansoni* immunophilin (AAB05213). The other *S. japonicum* sequences the motif was found in either lack homology to any other sequence or are homologous only to unknown proteins. The motif occurs between amino acids 40 and 105 in immunophilin, myosin light chain, thioredoxin, calreticulin, and calcium binding protein. It occurs outside this range, starting between amino acid 60 and 90, only in thioredoxin peroxidases (AAD17299, AAG15507, AAG15507, AAD40685, AAW25625 and AAW25436) and unknown *S. japonicum* proteins (AAW25402 and AAW24755).

Figure 15. MAST #2 high scoring sequences

Input: motif from MEME #2 and ~4000 full-length *S. mansoni* and *S. japonicum* sequences

New proteins are outlined in red.

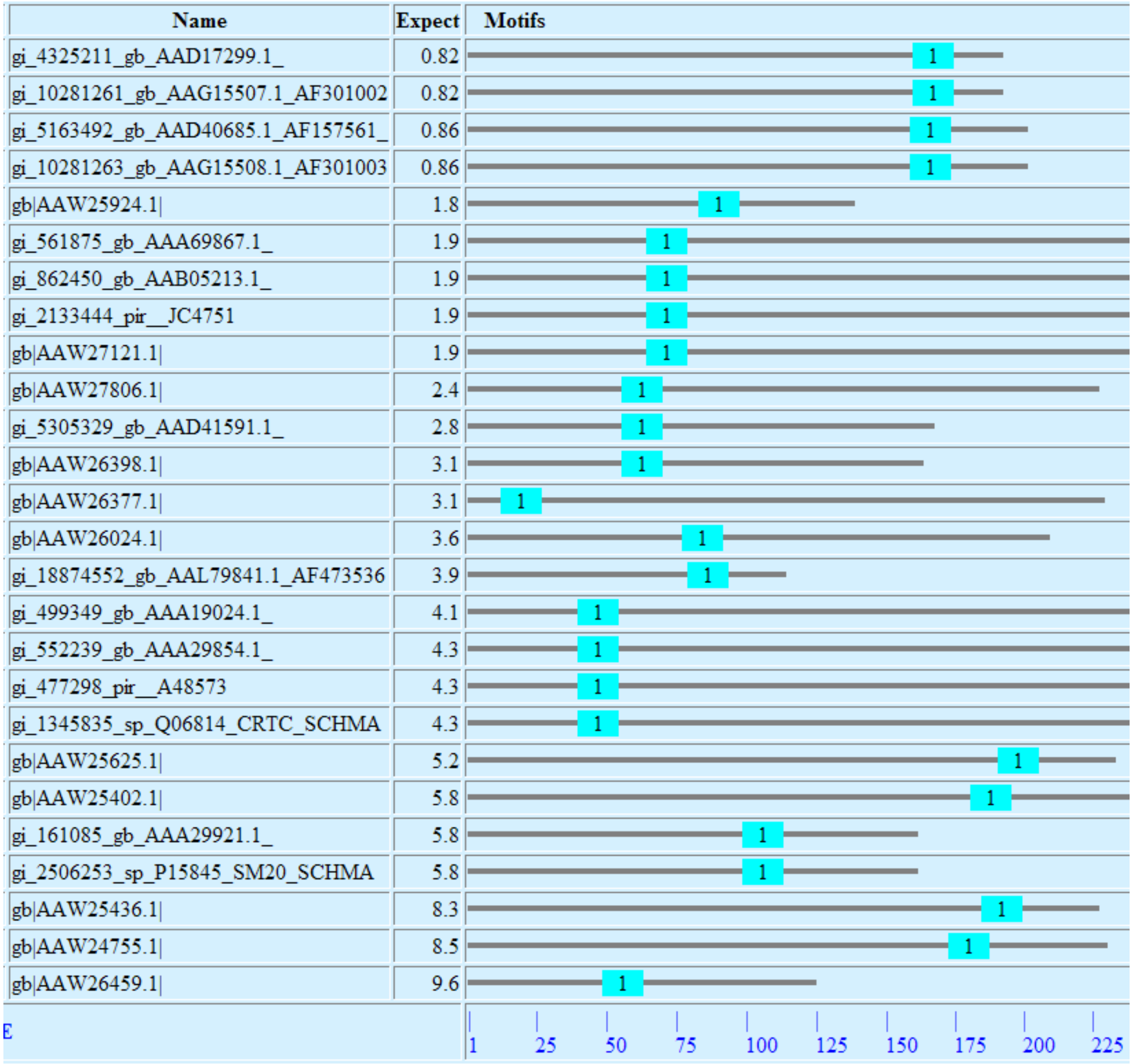


Sequence Name	**Description**	**E-value**	**Length**
gi_4325211_gb_AAD17299.1_	thioredoxin peroxidase [S...	0.82	185
gi_10281261_gb_AAG15507.1_AF301002	thioredoxin peroxidase 1 ...	0.82	185
gi_5163492_gb_AAD40685.1_AF157561_	thioredoxin peroxidase [S...	0.86	194
gi_10281263_gb_AAG15508.1_AF301003	thioredoxin peroxidase 2 ...	0.86	194
gb|AAW25924.1|	SJCHGC08819 protein [Schi...	1.8	131
gi_561875_gb_AAA69867.1_	immunophilin	1.9	430
gi_862450_gb_AAB05213.1_	immunophilin	1.9	430
gi_2133444_pir__JC4751	FK506-binding protein p50...	1.9	430
gb|AAW27121.1|	SJCHGC01391 protein [Schi...	1.9	431
gb|AAW27806.1|	SJCHGC08234 protein [Schi...	2.4	220
gi_5305329_gb_AAD41591.1_	myosin light chain [Schis...	2.8	160
gb|AAW26398.1|	SJCHGC01894 protein [Schi...	3.1	156
gb|AAW26377.1|	SJCHGC04329 protein [Schi...	3.1	222
gb|AAW26024.1|	SJCHGC04817 protein [Schi...	3.6	202
gi_18874552_gb_AAL79841.1_AF473536	thioredoxin [Schistosoma ...	3.9	106
gi_499349_gb_AAA19024.1_	calreticulin	4.1	373
gi_552239_gb_AAA29854.1_	antigen	4.3	393
gi_477298_pir__A48573	calreticulin autoantigen ...	4.3	393
gi_1345835_sp_Q06814_CRTC_SCHMA	Calreticulin precursor (S...	4.3	393
gb|AAW25625.1|	SJCHGC00794 protein [Schi...	5.2	226
gb|AAW25402.1|	SJCHGC09171 protein [Schi...	5.8	251
gi_161085_gb_AAA29921.1_	calcium binding protein [...	5.8	154
gi_2506253_sp_P15845_SM20_SCHMA	20 kDa calcium-binding pr...	5.8	154
gb|AAW25436.1|	SJCHGC01281 protein [Schi...	8.3	220
gb|AAW24755.1|	SJCHGC06903 protein [Schi...	8.5	223
gb|AAW26459.1|	SJCHGC04866 protein [Schi...	9.6	117

Figure 16. MAST #2 motif diagram

Input: motif from MEME #2 and ~4000 full-length *S. mansoni* and *S. japonicum* sequences

Motifs with significant position p-values (<.0001) are indicated by “1”.



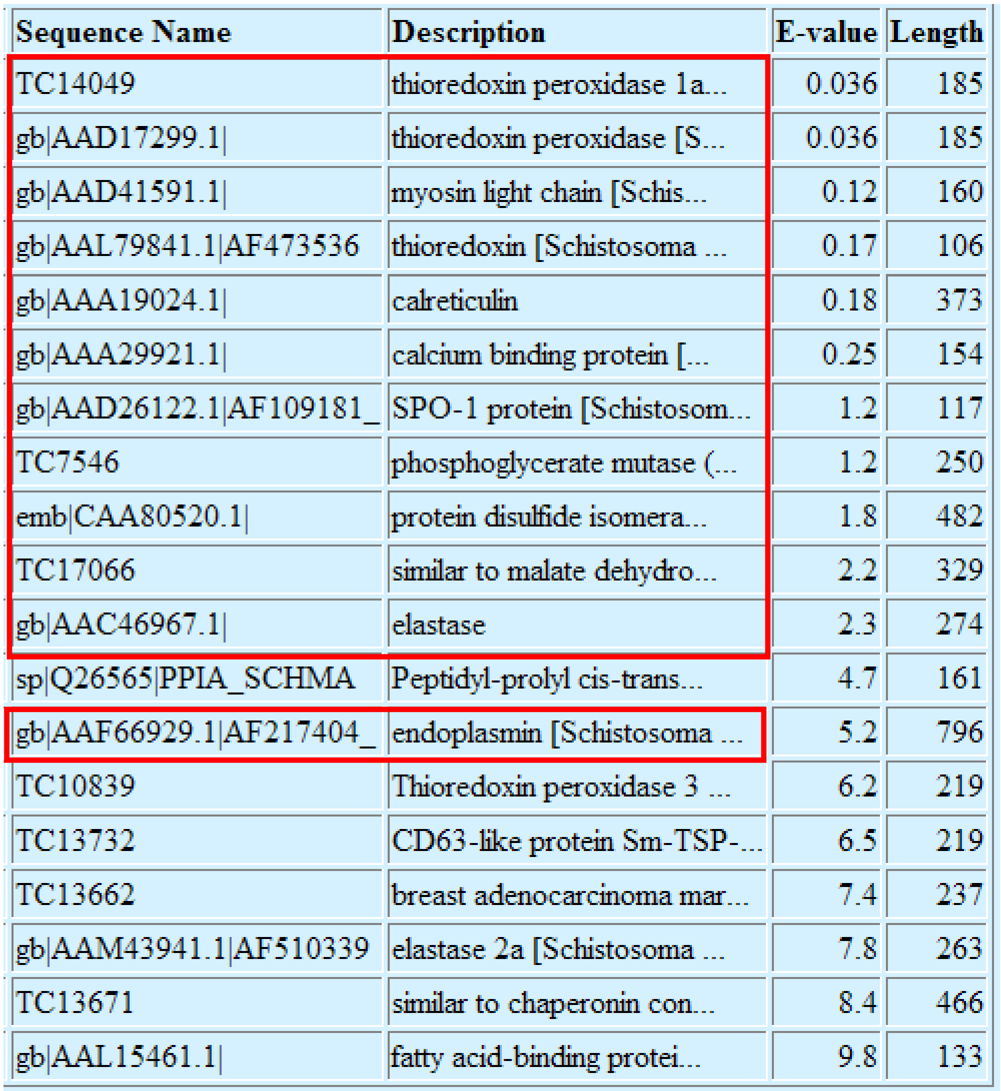
Name	Expect	Motifs
gi_4325211_gb_AAD17299.1_	0.82	1
gi_10281261_gb_AAG15507.1_AF301002	0.82	1
gi_5163492_gb_AAD40685.1_AF157561_	0.86	1
gi_10281263_gb_AAG15508.1_AF301003	0.86	1
gb|AAW25924.1|	1.8	1
gi_561875_gb_AAA69867.1_	1.9	1
gi_862450_gb_AAB05213.1_	1.9	1
gi_2133444_pir__JC4751	1.9	1
gb|AAW27121.1|	1.9	1
gb|AAW27806.1|	2.4	1
gi_5305329_gb_AAD41591.1_	2.8	1
gb|AAW26398.1|	3.1	1
gb|AAW26377.1|	3.1	1
gb|AAW26024.1|	3.6	1
gi_18874552_gb_AAL79841.1_AF473536	3.9	1
gi_499349_gb_AAA19024.1_	4.1	1
gi_552239_gb_AAA29854.1_	4.3	1
gi_477298_pir__A48573	4.3	1
gi_1345835_sp_Q06814_CRTC_SCHMA	4.3	1
gb|AAW25625.1|	5.2	1
gb|AAW25402.1|	5.8	1
gi_161085_gb_AAA29921.1_	5.8	1
gi_2506253_sp_P15845_SM20_SCHMA	5.8	1
gb|AAW25436.1|	8.3	1
gb|AAW24755.1|	8.5	1
gb|AAW26459.1|	9.6	1

Another MAST search, MAST #3, Figures 17 and 18, was executed using MEME motif #2 to search all original vesicle, secretion, and tegument proteins. Exact duplicate sequences were eliminated from this database resulting in a total of 130 proteins. Proteins with position p-values less than .0001 appear in the motif diagram in Figure 18. These include: thioredoxin, thioredoxin peroxidase 1a, thioredoxin peroxidase 2, myosin light chain, calreticulin, calcium binding protein, SPO-1 protein, PDI, endoplasmin, elastase 1b, phosphoglycerate mutase, and malate dehydrogenase. Elastase 1b, phosphoglycerate mutase, and malate dehydrogenase were new proteins, not discovered in previous MAST searches.

Figure 17. MAST #3 high scoring sequences

Input: motif from MEME #2 and all sample set proteins with redundancy removed (vesicles, secretions, tegument)

Motifs with significant position p-values (<.0001) are outlined in red.

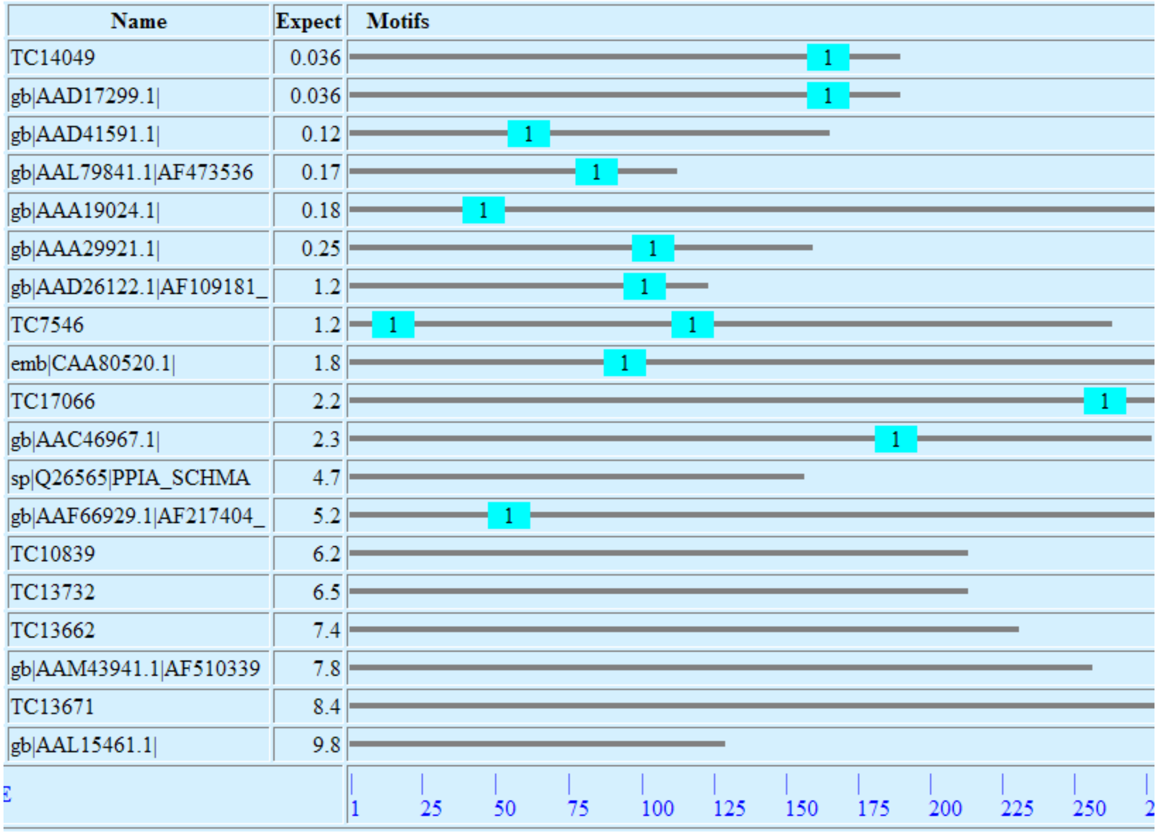


Sequence Name	Description	E-value	Length
TC14049	thioredoxin peroxidase 1a...	0.036	185
gb|AAD17299.1|	thioredoxin peroxidase [S...	0.036	185
gb|AAD41591.1|	myosin light chain [Schis...	0.12	160
gb|AAL79841.1|AF473536	thioredoxin [Schistosoma ...	0.17	106
gb|AAA19024.1|	calreticulin	0.18	373
gb|AAA29921.1|	calcium binding protein [...	0.25	154
gb|AAD26122.1|AF109181_	SPO-1 protein [Schistosom...	1.2	117
TC7546	phosphoglycerate mutase (...	1.2	250
emb|CAA80520.1|	protein disulfide isomera...	1.8	482
TC17066	similar to malate dehydro...	2.2	329
gb|AAC46967.1|	elastase	2.3	274
sp|Q26565|PPIA_SCHMA	Peptidyl-prolyl cis-trans...	4.7	161
gb|AAF66929.1|AF217404_	endoplasmin [Schistosoma ...	5.2	796
TC10839	Thioredoxin peroxidase 3 ...	6.2	219
TC13732	CD63-like protein Sm-TSP-...	6.5	219
TC13662	breast adenocarcinoma mar...	7.4	237
gb|AAM43941.1|AF510339	elastase 2a [Schistosoma ...	7.8	263
TC13671	similar to chaperonin con...	8.4	466
gb|AAL15461.1|	fatty acid-binding protei...	9.8	133

Figure 18. MAST #3 motif diagram

Input: motif from MEME #2 and all sample set proteins with redundancy removed (vesicles, secretions, tegument)

Motifs with significant position p-values (<.0001) indicated by “1”.



Name	Expect	Motifs
TC14049	0.036	1
gb|AAD17299.1|	0.036	1
gb|AAD41591.1|	0.12	1
gb|AAL79841.1|AF473536	0.17	1
gb|AAA19024.1|	0.18	1
gb|AAA29921.1|	0.25	1
gb|AAD26122.1|AF109181_	1.2	1
TC7546	1.2	1 1
emb|CAA80520.1|	1.8	1
TC17066	2.2	1
gb|AAC46967.1|	2.3	1
sp|Q26565|PPIA_SCHMA	4.7	
gb|AAF66929.1|AF217404_	5.2	1
TC10839	6.2	
TC13732	6.5	
TC13662	7.4	
gb|AAM43941.1|AF510339	7.8	
TC13671	8.4	
gb|AAL15461.1|	9.8	

Table 4. Sample proteins containing xKxGE motif

Based on MAST #3. Note that the motif is not conserved in SPO-1 protein shown in red and it was not carried forward into the rest of the analyses.

An asterisk (*) in the ‘SS?’ column indicates that SignalP did not predict a signal peptide but SIGCLEAVE did predict a signal peptide.

Protein	**V**	**S**	**T**	**protein length (aa)**	**SS?**	**SS cleavage site**	**motif**	**motif start**	**motif combined p-value**	**motif e-value**	**motif position p-value**
thioredoxin	AAL79841	AAL79841		106	yes*	37	IKNGE	80	1.01E-03	0.17	1.00E-05
thioredoxin peroxidase 1a			TC14049	185	no	n/a	EKHGE	162	2.11E-04	0.036	1.00E-06
thioredoxin peroxidase 2	AAD17299	AAD17299		185	no	n/a	EKHGE	162	2.11E-04	0.036	1.00E-06
myosin light chain	AAD41591			106	no	n/a	IKQGE	56	7.23E-04	0.12	5.00E-06
calreticulin	AAA19024			373	yes	17	EKQGE	40	1.05E-03	0.18	3.00E-06
calcium binding protein	AAA29921	AAA29921		154	no	n/a	NKRGE	100	1.49E-03	0.25	1.00E-05
SPO-1 protein	AAD26122	AAD26122		117	yes	17	EKRME	97	7.21E-03	1.2	6.00E-05
PDI	CAA80520			482	yes	18	LKHGE	90	1.04E-02	1.8	2.00E-05
elastase 1b	AAC46967	AAC46967		274	yes*	19	LKKGE	186	1.33E-02	2.3	5.00E-05
endoplasmin	AAF66929			796	yes	21	TKEGE	49	3.10E-02	5.2	4.00E-05
phosphoglycerate mutase	TC7546	TC7546		250	no	n/a	IRHGE, AKHGE	8, 91	7.33E-03	1.2	9.00E-05, 3.00E-05
malate dehydrogenase		TC17066		329	no	n/a	TKEGE	260	1.29E-02	2.2	4.00E-05

For each sample protein predicted by MAST to contain the proposed conserved motif, Table 4 (above) provides details about the motif prediction as well as signal peptide prediction, where applicable. Only three of the eleven proteins (27%) were predicted to contain a signal peptide by SignalP. These eleven proteins were further analyzed for signal peptides with SIGCLEAVE (Peter Rice, European Bioinformatics Institute, 1989). SIGCLEAVE uses the von Heijne method to distinguish signal peptides from non-signal peptides (95% accuracy) and to predict the cleavage site (75-80% accuracy). This analysis predicted that five of the eleven proteins had signal peptides: the same three predicted by SignalP and also thioredoxin and elastase 1b.

The SPO-1 protein does not actually contain the xKxGE motif. Starting in position 97 the amino acid sequence xKx**M**E exists (shown in red in Table 4). It is not likely that methionine can substitute for glycine. Other SPO-1 sequences were evaluated to ensure the difference was not due to sequencing error and they were found to also xKxME instead of xKxGE. Hence, SPO-1 was not considered to contain the motif and not included in further analyses. CLUSTALW analysis of the remaining eleven proteins and two additional public domain proteins found to contain the xKxGE motif was performed. The resulting multiple sequence alignment, which has gaps removed, is shown in Figure 19. Predicted signal peptides are underlined in red. Cleavage sites are at the C-terminal of the signal peptide. The downstream conserved motif predicted by MEME and MAST is underlined in blue for each protein. MEME/MAST does not provide cleavage site predictions.

The alignment shows that the motif is predicted to occur at positions 261-265 in malate dehydrogenase, a protein that is 329 amino acids long. Given the predicted position and the fact that the e-value of the motif in malate dehydrogenase was higher (2.2) than for any of the other eleven proteins, it is unlikely that the motif is not real in that protein. Phosphoglycerate mutase also has the motif in an unlikely position- amino acid 9-13. This does not leave room for a signal peptide to target the protein to the ER. (It was verified using BLAST that the N-terminal of the protein was included.) These two proteins were not considered in further analyses to have the motif of interest.

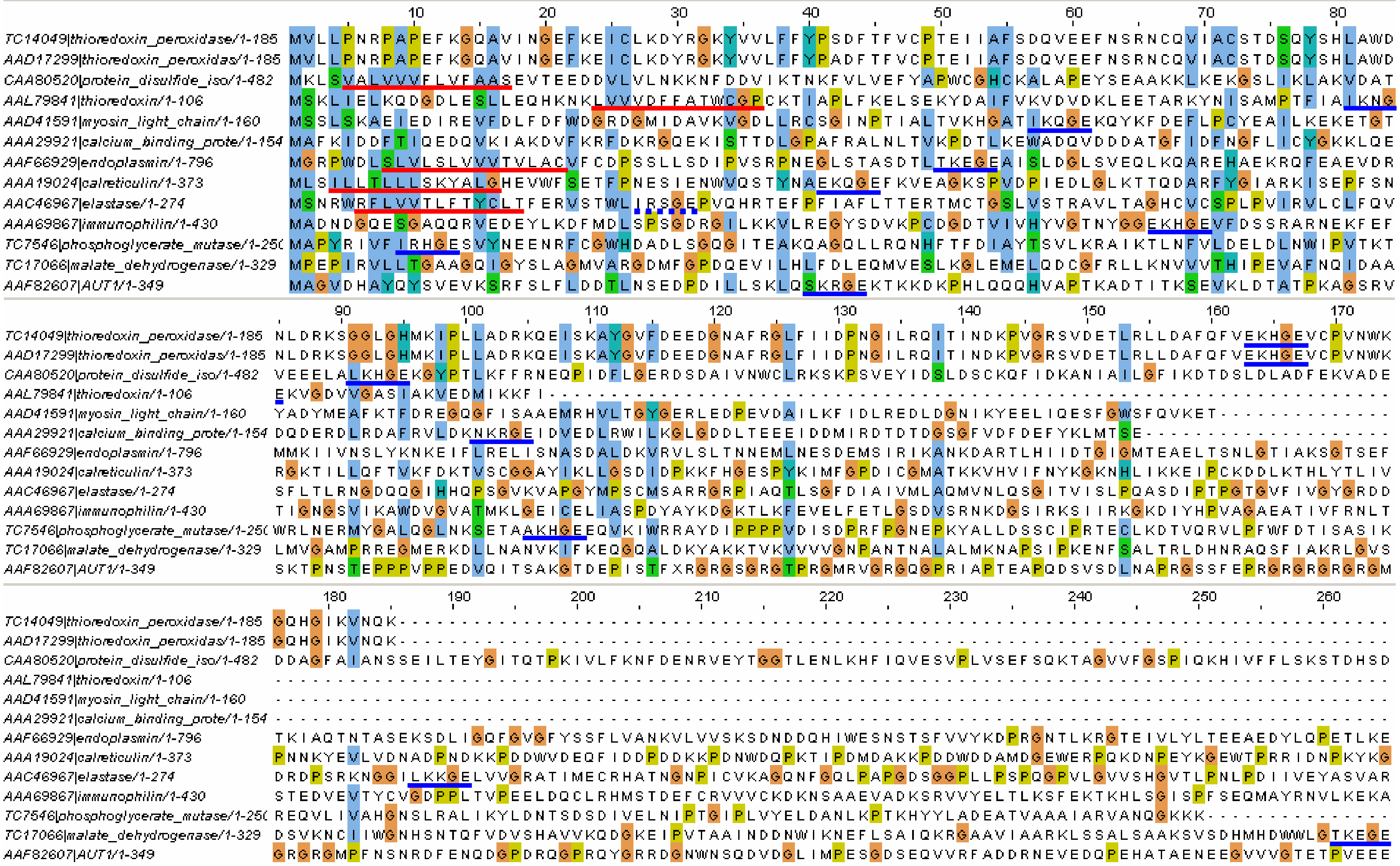

TC14049|thioredoxin_peroxidase/1-185
AAD17299|thioredoxin_peroxidas/1-185
CAA80520|protein_disulfide_iso/1-482
AAL79841|thioredoxin/1-106
AAD41591|myosin_light_chain/1-160
AAA29921|calcium_binding_prote/1-154
AAF66929|endoplasmin/1-796
AAA19024|calreticulin/1-373
AAC46967|elastase/1-274
AAA69867|immunophilin/1-430
TC7546|phosphoglycerate_mutase/1-25(
TC17066|malate_dehydrogenase/1-329
AAF82607|AUT1/1-349

Figure 19. Multiple sequence alignment of sample proteins with the predicted xKxGE motif (CLUSTALW alignment with gaps removed)

Elastase, of which there are many isoforms, is well-known to be secreted by *Schistosoma* during human invasion. The xKxGE motif was only found in one of the three isoforms identified in vesicles and/or secretions (elastase 1b) and it was found in an unlikely position- 186. Results from MEME #1 (page 65), however, show that in all three elastases contained in the sample proteins (AAC46967, A28941, and AAM43941), a motif of x**R**xGE exists starting two amino acids downstream of a signal sequence. Theorizing that the motif is a proprotein convertase recognition site and considering the fact that, at least in mammals, proprotein convertases (PCs) are known to cleave immediately after a consensus sequence that starts with either K or R (discussed on page 36), we can extrapolate that arginine (R) substitutes for lysine (K) in the second position of our proposed conserved motif in elastases. Figure 19 shows the motif IRSGE in elastase 1b (AAC46967) underlined in a dotted blue line starting in position 27, and the motif predicted by MEME/MAST underlined in a solid blue line starting in position 186.

Based on all MEME and MAST analyses, two additional *S. mansoni* proteins not contained in any of the sample groups contain the motif: immunophilin and AUT1. Immunophilin has 88% identity to an *S. japonicum* protein identified only as SJCHGC01391. It is not predicted to contain a signal peptide but has the conserved motif EKHGE starting at amino acid 66. Two *S. mansoni* AUT1 proteins exist in NCBI's nonredundant protein database, but only have homology to unidentified proteins. AUT1 was also predicted not to have a signal peptide. It has the conserved motif SKRGE starting at position 40.

If, based on the above findings, SPO-1, malate dehydrogenase, and phosphoglycerate mutase are eliminated from those proteins we consider to have the conserved motif, and all three elastases *are* considered to have the motif, a total of 11 sample proteins and 2 non-sample proteins are included, as summarized in Figure 20. 10 of 81 (12.3%) vesicle proteins, 6 of 53 (11.3%) secretion proteins, and only 1 of 43 (2.3%) tegument proteins contain the motif. As the diagram shows, all 6 proteins contained in secretions that have the motif are also contained in vesicles- none are unique to secretions. However, 4 of the 10 proteins in vesicles found to have the motif *are* unique to vesicles. Hence, the motif is enriched in the vesicles.

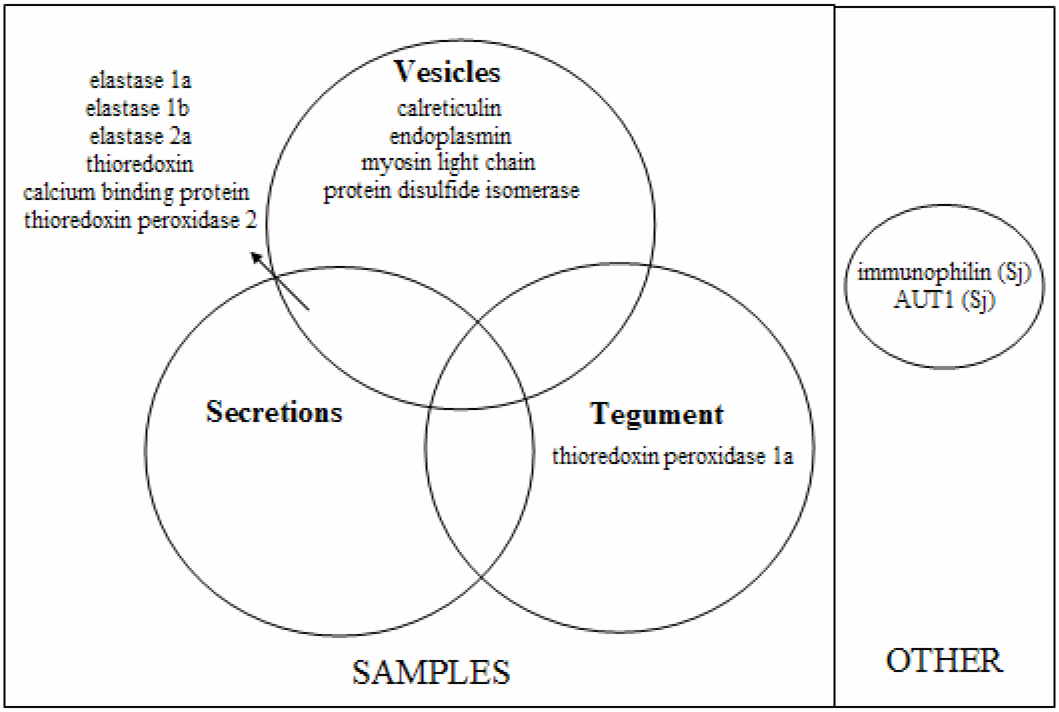
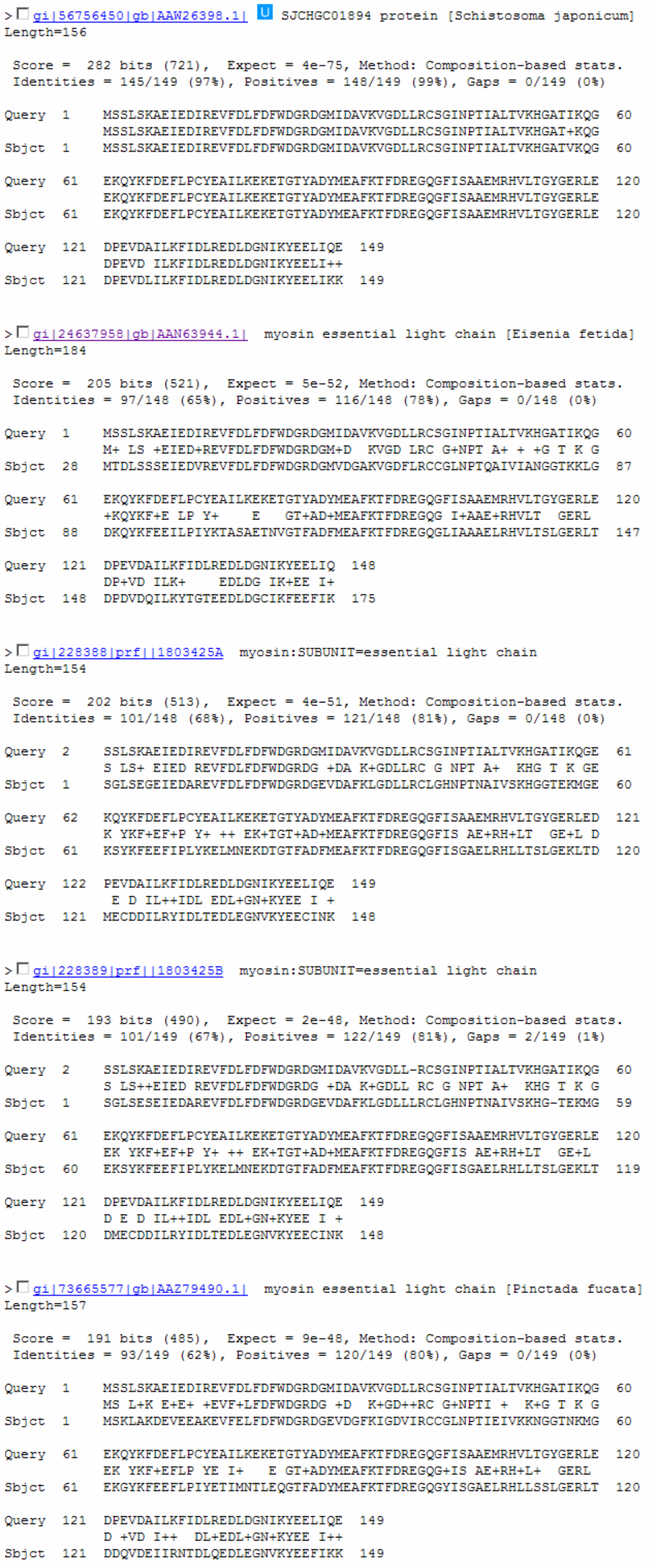


Figure 20. Distribution of proteins found to contain xKxGE motif.
Based on MAST #3

The sequences with the motif but without a predicted signal sequence were analyzed further to determine if the reason no signal sequence was predicted was because the N-terminus of the sequence was missing. A BLAST search against NCBI's nonredundant protein database was performed using each of these sequences and the starting positions of the query sequence and homologous sequences were compared. Myosin light chain, for example, was predicted not to contain a signal peptide. The top five BLAST hits for myosin light chain are shown in Figure 21. In four of the five hits the homology between the query sequence and subject sequence starts at position 1 or position 2. But in the second hit to myosin light chain in *Eisenia fetida* the homology starts at position 1 of the



```
>☐gi|56756450|gb|AAW26398.1| U SJCHGC01894 protein [Schistosoma japonicum]
Length=156

 Score =  282 bits (721),  Expect = 4e-75, Method: Composition-based stats.
 Identities = 145/149 (97%), Positives = 148/149 (99%), Gaps = 0/149 (0%)

Query  1    MSSLSKAEIEDIREVFDLFDFWDGRDGMIDAVKVGDLLRCSGINPTIALTVKHGATIKQG  60
            MSSLSKAEIEDIREVFDLFDFWDGRDGMIDAVKVGDLLRCSGINPTIALTVKHGAT+KQG
Sbjct  1    MSSLSKAEIEDIREVFDLFDFWDGRDGMIDAVKVGDLLRCSGINPTIALTVKHGATVKQG  60

Query  61   EKQYKFDEFLPCYEAILKEKETGTYADYMEAFKTFDREGQGFISAAEMRHVLTGYGERLE  120
            EKQYKFDEFLPCYEAILKEKETGTYADYMEAFKTFDREGQGFISAAEMRHVLTGYGERLE
Sbjct  61   EKQYKFDEFLPCYEAILKEKETGTYADYMEAFKTFDREGQGFISAAEMRHVLTGYGERLE  120

Query  121  DPEVDAILKFIDLREDLDGNIKYEELIQE  149
            DPEVD ILKFIDLREDLDGNIKYEELI++
Sbjct  121  DPEVDLILKFIDLREDLDGNIKYEELIKK  149


>☐gi|24637958|gb|AAN63944.1|  myosin essential light chain [Eisenia fetida]
Length=184

 Score =  205 bits (521),  Expect = 5e-52, Method: Composition-based stats.
 Identities = 97/148 (65%), Positives = 116/148 (78%), Gaps = 0/148 (0%)

Query  1    MSSLSKAEIEDIREVFDLFDFWDGRDGMIDAVKVGDLLRCSGINPTIALTVKHGATIKQG  60
            M+ LS +EIED+REVFDLFDFWDGRDGM+D  KVGD LRC G+NPT A+ + +G T K G
Sbjct  28   MTDLSSSEIEDVREVFDLFDFWDGRDGMVDGAKVGDFLRCCGLNPTQAIVIANGGTKKLG  87

Query  61   EKQYKFDEFLPCYEAILKEKETGTYADYMEAFKTFDREGQGFISAAEMRHVLTGYGERLE  120
            +KQYKF+E LP Y+    E   GT+AD+MEAFKTFDREGQG I+AAE+RHVLT  GERL
Sbjct  88   DKQYKFEEILPIYKTASAETNVGTFADFMEAFKTFDREGQGLIAAAELRHVLTSLGERLT  147

Query  121  DPEVDAILKFIDLREDLDGNIKYEELIQ  148
            DP+VD ILK+    EDLDG IK+EE I+
Sbjct  148  DPDVDQILKYTGTEEDLDGCIKFEEFIK  175


>☐gi|228388|prf||1803425A  myosin:SUBUNIT=essential light chain
Length=154

 Score =  202 bits (513),  Expect = 4e-51, Method: Composition-based stats.
 Identities = 101/148 (68%), Positives = 121/148 (81%), Gaps = 0/148 (0%)

Query  2    SSLSKAEIEDIREVFDLFDFWDGRDGMIDAVKVGDLLRCSGINPTIALTVKHGATIKQGE  61
            S LS+ EIED REVFDLFDFWDGRDG +DA K+GDLLRC G NPT A+  KHG T K GE
Sbjct  1    SGLSEGEIEDAREVFDLFDFWDGRDGEVDAFKLGDLLRCLGHNPTNAIVSKHGGTEKMGE  60

Query  62   KQYKFDEFLPCYEAILKEKETGTYADYMEAFKTFDREGQGFISAAEMRHVLTGYGERLED  121
            K YKF+EF+P Y+ ++ EK+TGT+AD+MEAFKTFDREGQGFIS AE+RH+LT  GE+L D
Sbjct  61   KSYKFEEFIPLYKELMNEKDTGTFADFMEAFKTFDREGQGFISGAELRHLLTSLGEKLTD  120

Query  122  PEVDAILKFIDLREDLDGNIKYEELIQE  149
             E D IL++IDL EDL+GN+KYEE I +
Sbjct  121  MECDDILRYIDLTEDLEGNVKYEECINK  148


>☐gi|228389|prf||1803425B  myosin:SUBUNIT=essential light chain
Length=154

 Score =  193 bits (490),  Expect = 2e-48, Method: Composition-based stats.
 Identities = 101/149 (67%), Positives = 122/149 (81%), Gaps = 2/149 (1%)

Query  2    SSLSKAEIEDIREVFDLFDFWDGRDGMIDAVKVGDLL-RCSGINPTIALTVKHGATIKQG  60
            S LS++EIED REVFDLFDFWDGRDG +DA K+GDLL RC G NPT A+  KHG T K G
Sbjct  1    SGLSESEIEDAREVFDLFDFWDGRDGEVDAFKLGDLLLRCLGHNPTNAIVSKHG-TEKMG  59

Query  61   EKQYKFDEFLPCYEAILKEKETGTYADYMEAFKTFDREGQGFISAAEMRHVLTGYGERLE  120
            EK YKF+EF+P Y+ ++ EK+TGT+AD+MEAFKTFDREGQGFIS AE+RH+LT  GE+L
Sbjct  60   EKSYKFEEFIPLYKELMNEKDTGTFADFMEAFKTFDREGQGFISGAELRHLLTSLGEKLT  119

Query  121  DPEVDAILKFIDLREDLDGNIKYEELIQE  149
            D E D IL++IDL EDL+GN+KYEE I +
Sbjct  120  DMECDDILRYIDLTEDLEGNVKYEECINK  148


>☐gi|73665577|gb|AAZ79490.1|  myosin essential light chain [Pinctada fucata]
Length=157

 Score =  191 bits (485),  Expect = 9e-48, Method: Composition-based stats.
 Identities = 93/149 (62%), Positives = 120/149 (80%), Gaps = 0/149 (0%)

Query  1    MSSLSKAEIEDIREVFDLFDFWDGRDGMIDAVKVGDLLRCSGINPTIALTVKHGATIKQG  60
            MS L+K E+E+ +EVF+LFDFWDGRDG +D  K+GD++RC G+NPTI +  K+G T K G
Sbjct  1    MSKLAKDEVEEAKEVFELFDFWDGRDGEVDGFKIGDVIRCCGLNPTIEIVKKNGGTNKMG  60

Query  61   EKQYKFDEFLPCYEAILKEKETGTYADYMEAFKTFDREGQGFISAAEMRHVLTGYGERLE  120
            EK YKF+EFLP YE I+   E GT+ADYMEAFKTFDREGQG+IS AE+RH+L+  GERL
Sbjct  61   EKGYKFEEFLPIYETIMNTLEQGTFADYMEAFKTFDREGQGYISGAELRHLLSSLGERLT  120

Query  121  DPEVDAILKFIDLREDLDGNIKYEELIQE  149
            D +VD I++  DL+EDL+GN+KYEE I++
Sbjct  121  DDQVDEIIRNTDLQEDLEGNVKYEEFIKK  149
```

Figure 21. Top five BLAST hits for myosin light chain AAD41591

subject sequence that had a different N-terminal portion than the query sequence, the subject sequence was submitted to SignalP for signal peptide prediction. But in each case, SignalP results still indicated the lack of a signal peptide.

The logo for the proposed conserved motif, below in Figure 22, shows the relative conservation of each amino acid in the motif. The second, fourth, and fifth residues are most conserved corresponding to lysine (basic), glycine (nonpolar), and glutamic acid (acidic). This corresponds to an enrichment of Lysine, Glutamic Acid, and Glycine residues in the vesicles over the control data set. Interestingly, the motif does have features in common with a conserved motif found in *P. falciparum*, shown in Figure 23. The first most conserved residue in both motifs is a basic amino acid: lysine in *Schistosoma* and arginine in *Plasmodium*. The last most conserved residue in both motifs is glutamic acid. The central most conserved amino acids differ across the two, although there are amino acids with nonpolar side chains in both: in *Schistosoma* glycine is conserved, and in *Plasmodium* lysine and alanine are conserved. One difference is that serine, which has an uncharged polar side chain, is somewhat conserved in *Plasmodium.*

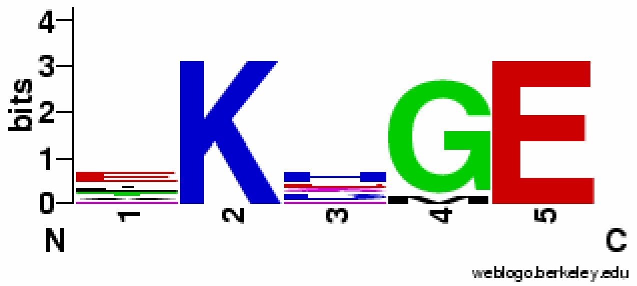


Figure 22. WebLogo of motif found in *S. mansoni* sample proteins
(Crooks, Hon et al. 2004)

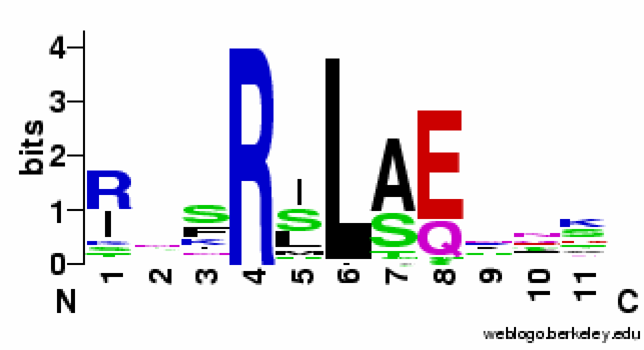


Figure 23. WebLogo of motif in *P. falciparum.*
(Crooks, Hon et al. 2004) Found across ~320 soluble and insoluble proteins that are exported across the PVM into the red blood cell (Hiller, Bhattacharjee *et al.* 2004).

Further analyses were performed in an effort to determine the possible functions of this motif. Based on the positions of the motifs and signal peptide prediction by SignalP, the motif is not a signal peptide. It was also determined to not be a glycosylation site, a protease site, a phosphorylation site, or an elastase recognition motif. BLAST searches determined that the set of proteins containing the motif do not have any significant homology to each other (Appendices C and D), hence MEME is not simply finding regions of homology across the proteins. Some propeptide convertases are known to cleave on the C-terminal side of sites composed of single or paired basic amino acids; it is possible that the motif is a propeptide that is cleaved on the C-terminal side of the second position of the motif, lysine or arginine (K/R).

Proteins containing the motif were submitted to ProP, a propeptide predictor based on neural networks (Duckert, Brunak et al. 2004). The general propeptide convertase-specific network was used, which was trained on 235 cleavage sites and 1072 negative sites from the SWISS-PROT protein database. It is important to note that the training sequences did not contain any *Schistosoma* sequences. The sensitivity of the network is 61.7, the specificity is 59.7, and the correlation coefficient is .60. Appendix E contains the positive results from this prediction method. 9 of the 14 proteins (64%) containing the motif predicted by MEME/MAST were predicted to have one or more propeptides by ProP: the three elastases, the two thioredoxin peroxidases, calcium binding protein, calreticulin, phosphoglycerate mutase, and AUT1. ProP only predicted the same motif/position for calcium binding protein that MEME/MAST predicted: NKRGE starting at position 100. Given the relatively low sensitivity and specificity of the ProP predictor, the MEME/MAST results were considered authoritative over the ProP results.

In order to determine how the proposed conserved motif fits into a secretion model, it is necessary to better understand the functions of the proteins that contain it. All sample proteins were functionally annotated with protein domains and GO terms using InterProScan. Methods were developed for transforming the data into matrices of protein domain and GO annotations versus sequence identifiers/descriptions, as explained under *Protein Domain/Gene Ontology Annotation and Statistical Analysis*. Smaller matrices containing the subset of sample proteins containing the proposed conserved motif were extracted from the larger matrices and are shown in Tables 5 and 6. These matrices provide a quick visualization of protein domains and GO annotations that the proteins do and do not have in common. They also give a starting point for more detailed analysis.

The domain matrix in Table 5 shows that several of the proteins have domains in common: 1) both the calcium binding protein and myosin light chain have an EF-Hand motif, 2) both thioredoxin and PDI have a thioredoxin-related domain, and 4) thioredoxin peroxidase 1a, thioredoxin peroxidase 2, thioredoxin, and PDI have a thioredoxin-like fold. The GO annotation matrix in Table 6 shows the following: both endoplasmin and calreticulin are involved in protein folding; calcium binding protein, myosin, and calreticulin are all involved in calcium ion binding; thioredoxin and PDI are involved in electron transport activity; and the elastases are involved in peptidase activity.

AUT1 (AAF82607, not included in Tables 5 or 6) contains a Hyaluronan/mRNA binding protein domain (IPR006861), which includes a family of proteins that have been shown to bind hyaluronan, a glucosaminoglycan. Immunophilin (AAA67867, not included in tables 5 or 6) contains a peptidylprolyl isomerase domain (IPR001179) as well as three tetratricopeptide-related domains (IPR001440, IPR11990, IPR13026). The peptidylprolyl isomerase domain is involved in accelerating protein folding by catalyzing the cis-trans isomerization of proline imidic peptide bonds in oligopeptides. The tetratricopeptide-related domains mediate protein-protein interactions and the assembly of multiprotein complexes. They are involved in a variety of biological processes including peroxisomal protein transport and protein folding.

Interestingly, endoplasmin, calreticulin, and PDI have historically all been known to reside in the ER lumen. Such ER-residing proteins are usually distinguished by the presence of the sequence KDEL at the C-terminal. This sequence does indeed exist at the C-terminal of PDI. In endoplasmin the C-terminal sequence is KNEL and in calreticulin the C-terminal sequence is HDEL. It has been shown, however, that PDI and endoplasmin can be secreted via the normal secretory pathway even when they contain an ER-retention signal (Takemoto, Yoshimori et al. 1992) and PDI has also found to be more highly concentrated in secretory cells (Edman, Ellis et al. 1985). It has been found that the active site of PDI closely resembles that of the redox protein thioredoxin and, therefore, may catalyze disulphide bond formation (Freedman R. B., Hirst T. R. et al. 1994).

To determine the relatedness of proteins within a sample group, a phylogenetic tree was also constructed for each sample group using CLUSTALW as explained under *Homology Analysis*. The complete phylogenetic trees are shown on pages 104-106. Proteins were assigned to phylogenetic groups as labeled on the trees according to the natural groups they would fall into if the tree was cut just below the main branch. The tables on pages 107-115 provide details about the functions related to the proteins contained in each phylogenetic group for each sample set.

Phylogenetic groupings, domains and GO terms are summarized in Table 7 for the vesicle and secretion proteins containing the proposed conserved motif. It shows that all of the proteins except the elastases belong to the same major phylogenetic group, 1. Five of the proteins belong to 1a, one belongs to 1d, one belongs to 1g, and the elastases belong to group 2. Phylogenetic group 1a is related to unfolded protein binding, calcium ion binding, motor activity, and electron transporter activity. Group 1d is related to electron transporter activity, glycolysis, and gluconeogenesis. Group 1g is related to a diverse set of functions including glycolysis, carbohydrate metabolism, protein polymerization, and protein folding. Group 2 corresponds to peptidase activity.

Considering all detailed annotation data together it is possible to theorize about the functions of the proteins containing the xKxGE motif. It was found in proteins related to calcium binding and regulation, protein folding, and protease activity. Calreticulin is implicated in a large and diverse number of functions. One of its key functions in *Schistosoma* cercaria may be to bind to the clathrin coat of the vesicles while still in the Golgi apparatus. Perhaps it also plays a role in the uptake of other proteins into the vesicles. Once inside the vesicles calreticulin likely plays a role in protein folding along with several other proteins. Myosin light chain and calcium binding protein are likely involved in motor activity that propels the vesicles out of the acetabular glands and guides them to their target epidermal cells. The proteins with calcium binding domains, which include calcium binding protein, myosin, and calreticulin, are likely to be important in the initial binding to target epidermal cells. PDI, endoplasmin, calreticulin, and the two proteins not in the samples but found in *S. japonicum*, immunophilin and AUT1, all have chaperone activity, which promotes proper protein folding to prevent protein aggregation and degradation by proteases. Though AUT1 was not found in the sample proteins, domain analysis indicates it has functions in common with calreticulin, and the proposed motif occurs in the same position in these two proteins. Thioredoxin and PDI, as discussed, are likely to function as catalysts for disulfide bond formation. The thioredoxin peroxidases are involved in oxidation of thioredoxin and have antioxidant activity which likely helps protect the secreted proteins from damage during human host invasion. Finally, the elastases have peptidase activity which is directly related to breaking down various layers of human host epidermal tissues.

Table 5. Proteins containing predicted conserved motif vs. protein domains

0 indicates the protein does not contain the corresponding domain, 1 indicates the protein does contain the corresponding domain.

accession	description	Vesicles, Secretions, or Tegument	IPR013078	IPR005952	IPR002048	IPR011992	IPR012336	IPR012335	IPR000866	IPR006662	IPR000886	IPR001314	IPR001254	IPR009003	IPR008256	IPR009079	IPR001404	IPR003299	IPR001580	IPR013320	IPR008985	IPR009169
			Phosphoglycerate mutase	Phosphoglycerate mutase 1	Calcium-binding EF-hand	EF-Hand type	Thioredoxin-like fold	Thioredoxin fold	Alkyl hydroperoxide reductase/ Thiol specific antioxidant/ Mal allergen	Thioredoxin	Endoplasmic reticulum targeting sequence	Peptidase S1A, chymotrypsin	Peptidase S1 and S6, chymotrypsin/Hap	Peptidase, trypsin-like serine and cysteine	Peptidase S1B, glutamyl endopeptidase I	Four-helical cytokine	Heat shock protein Hsp90	Flagellar calcium-binding protein calflagin	Calreticulin/calnexin	Concanavalin A-like lectin/glucanase, subgroup	Concanavalin A-like lectin/glucanase	Calreticulin
TC14049	thioredoxin peroxidase 1a	T	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
AAD17299	thioredoxin peroxidase 2	V, S	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
AAL79841	thioredoxin	V, S	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
AAA29921	calcium binding protein	V, S	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
AAC46967	elastase 1b	V, S	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0
A28942	elastase 1a	V, S	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0
AAM43941	elastase 2a	V, S	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0
AAD41591	myosin light chain	V	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AAF66929	endoplasmin	V	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1	0	0	0	0	0
AAA19024	calreticulin	V	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1
CAA80520	protein disulfide isomerase	V	0	0	0	0	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0

Table 6. Sample sequences containing predicted conserved motif vs. GO terms

0 indicates the protein does not contain the corresponding GO term, 1 indicates the protein does contain the corresponding GO term.

Note that thioredoxin peroxidase 1a and 2 do not have any corresponding GO terms.

accession	description	vesicles, secretions, or tegument	GO:0016868	GO:0051082	GO:0005509	GO:0005489	GO:0004252	GO:0008236	GO:0001539	GO:0006508	GO:0006118	GO:0006457	GO:0006096	GO:0009288	GO:0005783
			Molecular Function: intramolecular transferase activity	Molecular Function: unfolded protein binding	Molecular Function: calcium ion binding	Molecular Function: electron transporter activity	Molecular Function: serine-type endopeptidase activity	Molecular Function: serine-type peptidase activity	Biological Process: ciliary or flagellar motility	Biological Process: proteolysis	Biological Process: electron transport	Biological Process: protein folding	Biological Process: glycolysis	Cellular Component: flagellum	Cellular Component: endoplasmic reticulum
AAL79841	thioredoxin	V, S	0	0	0	1	0	0	0	0	1	0	0	0	0
AAA29921	calcium binding protein	V, S	0	0	1	0	0	0	1	0	0	0	0	1	0
AAC46967	elastase 1b	V, S	0	0	0	0	1	0	0	1	0	0	0	0	0
A28942	elastase 1a	V, S	0	0	0	0	1	0	0	1	0	0	0	0	0
AAM43941	elastase 2a	V, S	0	0	0	0	1	1	0	1	0	0	0	0	0
AAD41591	myosin light chain	V	0	0	1	0	0	0	0	0	0	0	0	0	0
AAF66929	endoplasmin	V	0	1	0	0	0	0	0	0	0	1	0	0	0
AAA19024	calreticulin	V	0	1	1	0	0	0	0	0	0	1	0	0	1
CAA80520	protein disulfide isomerase	V	0	0	0	1	0	0	0	0	1	0	0	0	0

Table 7. Summary of motif-containing vesicle and secretion proteins.

Description	Accession number	Phylogenetic Group	Sample Group (V= Vesicles, S= Secretions)	Protein domains	GO terms
Calcium-binding protein	AAA29921	1a	V, S	EF-Hand; calcium-binding; flagellar calcium-binding protein	calcium ion binding, ciliary or flagellar motility
Thioredoxin peroxidase 2	AAD17299	1a	V, S	Thioredoxin-like fold; alkyl hydroperoxide reductase; Thiol specific antioxidant	protein folding
Myosin light chain	AAD41591	1a	V	EF-Hand	calcium ion binding
Endoplasmin	AAF66929	1a	V	Four-helical cytokine; Heat shock protein Hsp90	unfolded protein binding; protein folding
Protein-disulfide isomerase homolog	CAA80520	1a	V	Thioredoxin fold; thioredoxin related	electron transport
Thioredoxin	AAL79841	1d	V, S	Thioredoxin fold; thioredoxin related	electron transport
Calreticulin	AAA19024	1g	V	calreticulin/calnexin; Concanavalin A-like lectin/glucanase, subgroup; Concanavalin A-like lectin/glucanase	unfolded protein binding; protein folding; calcium ion binding
Elastase (elastase 1b)	AAC46967	2	V, S	Peptidase S1A, chymotrypsin; Peptidase S1 and S6, chymotrypsin/Hap; Peptidase, trypsin-like serine and cysteine	serine-type endopeptidase activity; proteolysis
Pancreatic elastase precursor (elastase 1a)	A28942	2	V, S	Peptidase S1A, chymotrypsin; Peptidase S1 and S6, chymotrypsin/Hap; Peptidase, trypsin-like serine and cysteine	serine-type endopeptidase activity; proteolysis
Elastase 2a	AAM43941	2	V, S	Peptidase S1A, chymotrypsin; Peptidase S1 and S6, chymotrypsin/Hap; Peptidase, trypsin-like serine and cysteine; Peptidase S1B, glutamyl endopeptidase I	serine-type endopeptidase activity; proteolysis; serine-type peptidase activity

Homology Analysis

To determine the level of identity within and across sample sets, BLAST searches were performed using each sample set as the query against either itself or a different sample set. The blastp program was used with the default settings. A Perl script was used to parse all hits where the total length of the high scoring pair (HSP) was greater than 100, the hit significance was better than 1 x 10^{-05}, and the percent identity was greater than 50%. Results are shown in Appendices D (homology within sample sets) and E (homology across sample sets).

None of the sample sets has a high level of identity within it. Within the vesicle proteins there are three elastases with 80% or greater homology to each other and three other proteins with 55% - 66% homology to each other, including tropomyosin, thioredoxin peroxidase, and a 14-3-3 protein homolog. Within the secretion proteins, the same three elastases exist and also two peptidyl-prolyl cis-trans isomerase proteins with 57% homology to each other. Within the tegument proteins there are two unidentified proteins with 74% homology to each other.

Two of fifty three secretion proteins showed ~99% homology to tegument proteins; these were thioredoxin peroxidase and ATP dephosphohydrolase I. Similarly, two of eighty one vesicle proteins had high homology to tegument proteins- thioredoxin peroxidase and fimbrin. Two other vesicle proteins, a major egg antigen and an unidentified protein, showed about 50% homology to two unidentified tegument proteins. These intersections are illustrated in Figure 24.

Considerably higher homology was found between vesicle and secretion proteins, as expected. Of the 53 secretion proteins, 41, or 77% were identical to vesicle proteins. Homology between the other proteins was not high. There are several proteins unique to either secretions or vesicles. The protein domains and GO terms that correspond to each set of proteins that are unique to a specific sample set are shown in Figures 25-29.

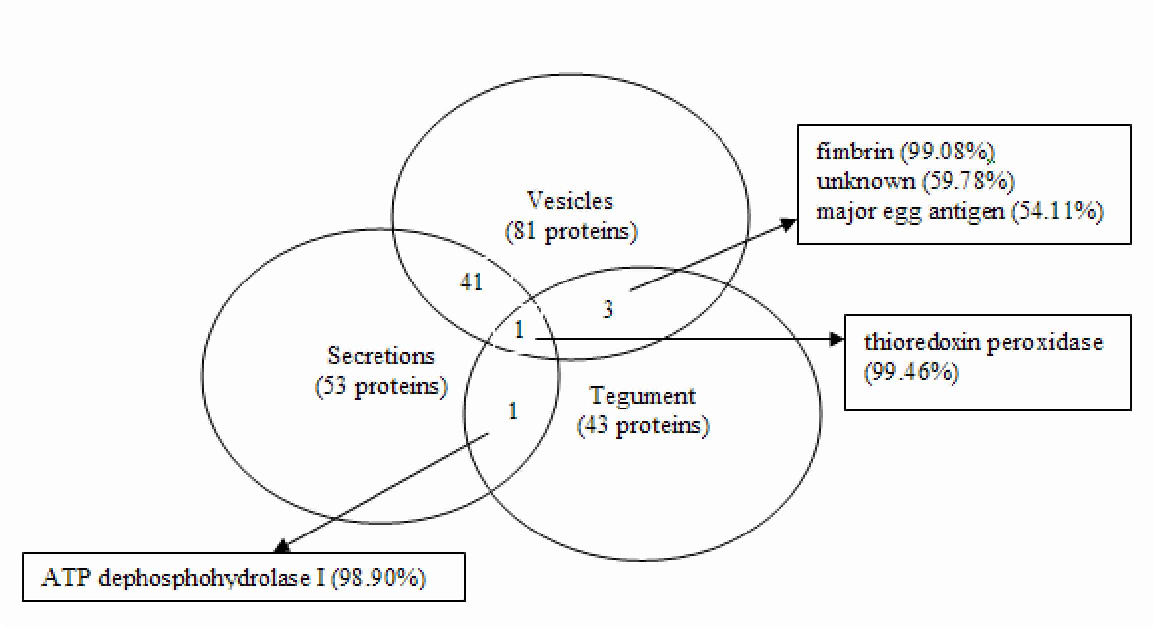


Figure 24. Venn diagram of similarity across sample proteins
Numbers in parentheses represent the percent similarity between the proteins across sample types.

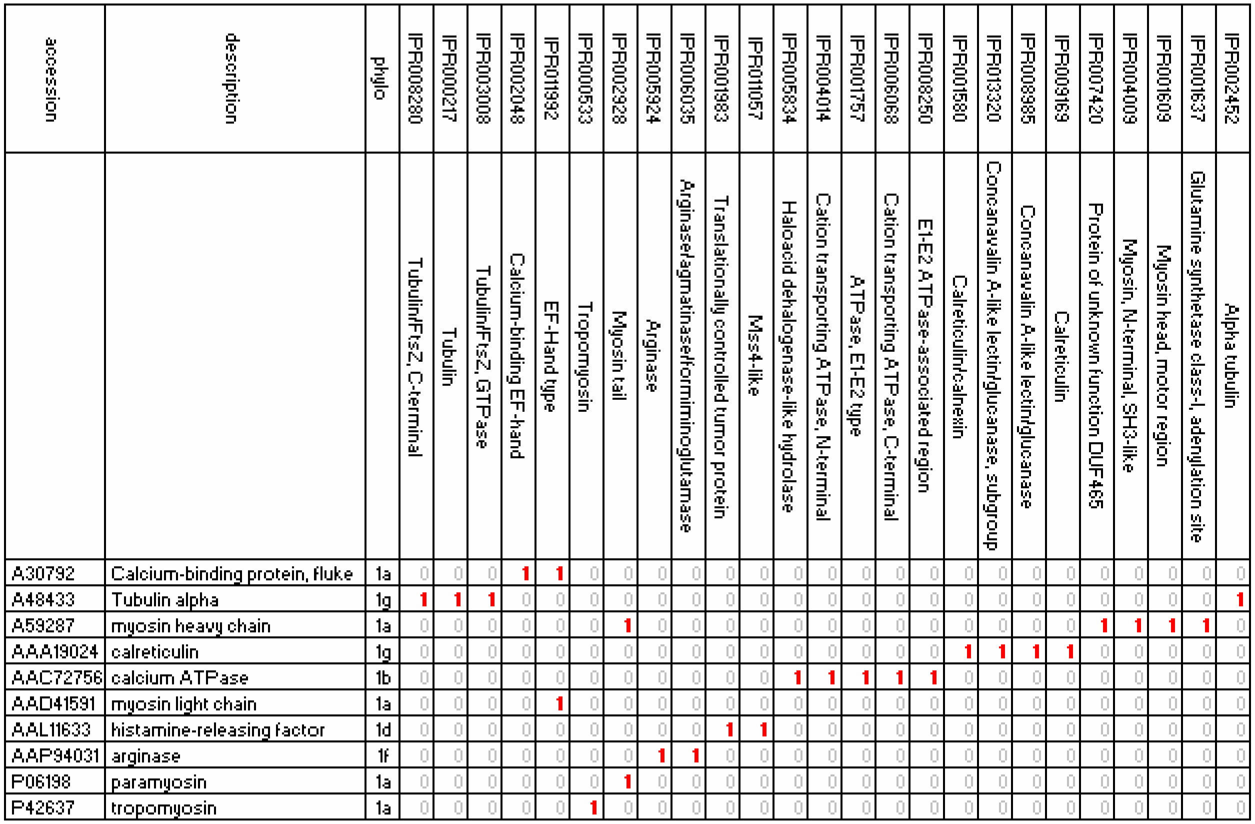
Proteins unique to vesicles included: calcium-binding protein (fluke), tubulin, myosin heavy chain, myosin light chain, paramyosin, tropomyosin, calreticulin, calcium ATPase, translationally-controlled tumor protein (TCTP)/histamine-releasing factor, and arginase. Calcium-binding protein, tubulin, myosin heavy chain, myosin light chain, paramyosin, and tropomyosin are all related to motor activity and microtubule-based movement and are likely used in early stages of human host invasion to guide the vesicles to their target. The role of TCTP/histamine-releasing factor is unknown, but it has been shown to bind to tubulin and has a high affinity for calcium. It is also the binding target for the anti-malarial compound artemisinin. The reason that so many of these proteins are uniquely found in the mechanically-induced sample compared to the lipid-induced sample may be related to the method used to collect the lipid-induced secretions. Cercarial bodies and heads were separated from lipid-induced secretions by centrifugation (Knudsen, Medzihradszky et al. 2005). This process may have also separated out the aggregated vesicular membranes, where proteins related to vesicular movement were attached. Mechanical induction secretions were separated by scraping vesicles from the surface of the plate, which retained the vesicular membrane in the sample. Calreticulin is implicated in protein folding and may be degraded in the lipid-induced secretions. Both calcium ATPase and arginase are metabolic enzymes that are responsible for extrusion of calcium and nitrogen respectively and, as such, would not be expected to be contained in the vesicles. Calcium ATPase contains a transmembrane domain.

Secretion proteins that did not show homology to either of the other two sample sets include: Cu,Zn-superoxide dismutase, ferritin-2 heavy chain, 6-phosphofructokinase, fatty acid-binding protein, calpain, cysteine protease inhibitor, and dynein light chain. Cu,Zn-superoxide dismutase, ferritin-2 heavy chain, 6-phosphofructokinase and fatty acid-binding protein are all soluble proteins. Calpain is an intracellular cysteine protease that is regulated by calcium. Its inhibitor is also unique to the secretions sample. As raised by Knudsen *et al.*, one possible reason for the presence of soluble proteins in the vesicle secretions is that the secretions do not just contain vesicle proteins, but also proteins contained in the cytosol of the acetabular glands (Knudsen, Medzihradszky et al. 2005). With the exception of dynein light chain, this could explain the presence of all of the proteins unique to the lipid-induced sample.

The proteins found in the tegument were quite different than those found in the vesicles and secretions, including such proteins as amidase, dysferlin, aquaporin, and many unidentified proteins. These proteins, not surprisingly, are all related to transport.

To determine the evolutionary relationships between the proteins contained in each sample set, a phylogenetic tree was created for each, as shown in Figures 30-32. The trees were generated using the alignment files resulting from the CLUSTALW alignment of each sample set. Proteins were then assigned to phylogenetic groups as labeled on the trees according to the natural groups they would fall into if the tree was cut just below the main branch. Table 8 provides full details about the protein domains and GO functions related to the proteins contained in each phylogenetic group for each sample set.

The vesicle, secretion, and tegument proteins seem to have four, six, and three major phylogenetic groupings respectively. In the vesicles there are four major phylogenetic groups. Group 1 is very large and represents 55 proteins with a diverse set of functions including calcium-binding protein and microtubule and motor-based movement proteins. Group two has 17 members representing mainly chaperones and the elastases. Groups three and four are very small with glycolysis and metabolism represented. In the secretions there are three major groups. The chaperones, calcium-binding proteins, and elastases group together in the same major phylogenetic group, group one. Group two contains glycolytic and metabolic enzymes as well as two antigenic proteins. Group three relates to proton, electron, and iron transport as well as microtubule-based processes. In the tegument there are also 3 major groups. Group one is the largest group representing a diverse set of processes including transport, protein modification, biosynthesis and targeting, and metabolism. Group two contains only proteins related to transport and group three contains proteins related to exocytosis/vesicle docking and nucleotide metabolism.



accession	description	phylo	IPR008280	IPR000217	IPR003008	IPR002048	IPR011992	IPR000533	IPR002928	IPR005924	IPR006035	IPR001983	IPR011057	IPR005834	IPR004014	IPR001757	IPR006068	IPR008250	IPR001580	IPR013320	IPR008985	IPR009169	IPR007420	IPR004009	IPR001609	IPR001637	IPR002452
			Tubulin/FtsZ, C-terminal	Tubulin	Tubulin/FtsZ, GTPase	Calcium-binding EF-hand	EF-Hand type	Tropomyosin	Myosin tail	Arginase	Arginase/agmatinase/formiminoglutamase	Translationally controlled tumor protein	Mss4-like	Haloacid dehalogenase-like hydrolase	Cation transporting ATPase, N-terminal	ATPase, E1-E2 type	Cation transporting ATPase, C-terminal	E1-E2 ATPase-associated region	Calreticulin/calnexin	Concanavalin A-like lectin/glucanase, subgroup	Concanavalin A-like lectin/glucanase	Calreticulin	Protein of unknown function DUF465	Myosin, N-terminal, SH3-like	Myosin head, motor region	Glutamine synthetase class-I, adenylation site	Alpha tubulin
A30792	Calcium-binding protein, fluke	1a	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A48433	Tubulin alpha	1g	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
A59287	myosin heavy chain	1a	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0
AAA19024	calreticulin	1g	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0
AAC72756	calcium ATPase	1b	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0	0	0	0	0
AAD41591	myosin light chain	1a	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AAL11633	histamine-releasing factor	1d	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AAP94031	arginase	1f	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
P06198	paramyosin	1a	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
P42637	tropomyosin	1a	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 25. Protein domains corresponding to vesicle-specific proteins

accession	description	phylo	GO:0005524	GO:0005525	GO:0005737	GO:0003924	GO:0043234	GO:0051258	GO:0005198	GO:0005874	GO:0007018	GO:0006457	GO:0051082	GO:0016020	GO:0005509	GO:0016820	GO:0008152	GO:0003774	GO:0016459	GO:0004053	GO:0006527	GO:0003824	GO:0005554	GO:0006812	GO:0015662	GO:0005783	GO:0004356	GO:0006807
			Molecular Function: ATP binding	Molecular Function: GTP binding	Cellular Component: cytoplasm	Molecular Function: GTPase activity	Cellular Component: protein complex	Biological Process: protein polymerization	Molecular Function: structural molecule activity	Cellular Component: microtubule	Biological Process: microtubule-based movement	Biological Process: protein folding	Molecular Function: unfolded protein binding	Cellular Component: membrane	Molecular Function: calcium ion binding	Molecular Function: hydrolase activity	Biological Process: metabolism	Molecular Function: motor activity	Cellular Component: myosin	Molecular Function: arginase activity	Biological Process: arginine catabolism	Molecular Function: catalytic activity	Molecular Function: molecular function unknown	Biological Process: cation transport	Molecular Function: ATPase activity	Cellular Component: endoplasmic reticulum	Molecular Function: glutamate-ammonia ligase activity	Biological Process: nitrogen compound metabolism
AAP94031	arginase	1f	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0
AAC72756	calcium ATPase	1b	1	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0	0	0	1	0	1	1	0	0	0
A30792	Calcium-binding protein, fluke	1a	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
AAA19024	calreticulin	1g	0	0	0	0	0	0	0	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0
AAL11633	histamine-releasing factor	1d	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
A59287	myosin heavy chain	1a	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	1	1
AAD41591	myosin light chain	1a	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
P06198	paramyosin	1a	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0
A48433	Tubulin alpha	1g	0	1	0	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

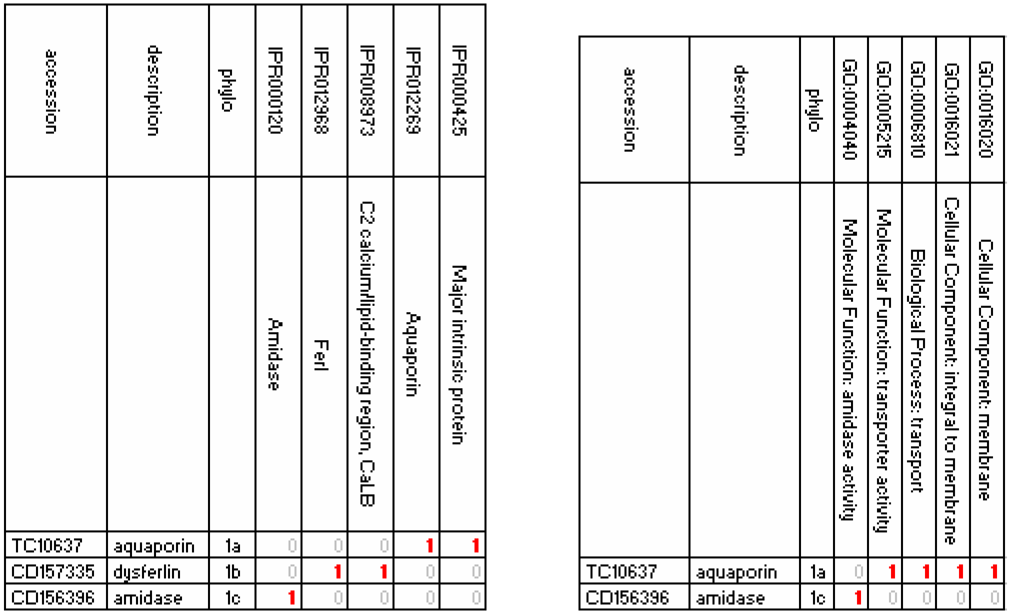
Figure 26. GO terms corresponding to vesicle-specific proteins.

accession	description	phylo	IPR001300 Peptidase C2, calpain	IPR011992 EF-Hand type	IPR000169 Peptidase, cysteine peptidase active site	IPR001424 Superoxide dismutase, copper/zinc binding	IPR000463 Cytosolic fatty-acid binding	IPR012674 Calycin	IPR000566 Lipocalin-related protein and Bos/Can/Equ allergen	IPR011038 Calycin-like	IPR000407 Nucleoside phosphatase GDA1/CD39	IPR000010 Proteinase inhibitor I25, cystatin	IPR001713 Proteinase inhibitor I25A, stefin A	IPR001519 Ferritin	IPR012347 Ferritin-related	IPR008331 Ferritin and Dps	IPR009078 Ferritin/ribonucleotide reductase-like	IPR000023 Phosphofructokinase	IPR009161 6-phosphofructokinase, eukaryotic type	IPR000941 Enolase	IPR001372 Dynein light chain, type 1
A39343	Calpain (EC 3.4.22.17) large chain (Sm)	1a	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AAL15461	Fatty acid-binding protein Sm14 (Sm)	1a	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0
AAC14467	Cu,Zn-superoxide dismutase (Sm)	1b	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AAQ16180	Cysteine protease inhibitor (Sm)	1c	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0
Q27778	6-Phosphofructokinase (Sm)	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0
P25319	Ferritin-2 heavy chain (Sm)	3	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0
Q94748	Probable dynein light chain (SM10) (T-cell-stimulating antigen SM10)	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Figure 27. Protein domains corresponding to secretion-specific proteins

accession	description	phylo	GO:0005622	GO:0004785	GO:0006801	GO:0046872	GO:0005488	GO:0006810	GO:0008289	GO:0004869	GO:0004866	GO:0006096	GO:0006826	GO:0006879	GO:0008199	GO:0003872	GO:0005945	GO:0005737	GO:0003777	GO:0005875	GO:0007017
			Cellular Component: intracellular	Molecular Function: copper, zinc superoxide dismutase activity	Biological Process: superoxide metabolism	Molecular Function: metal ion binding	Molecular Function: binding	Biological Process: transport	Molecular Function: lipid binding	Molecular Function: cysteine protease inhibitor activity	Molecular Function: endopeptidase inhibitor activity	Biological Process: glycolysis	Biological Process: iron ion transport	Biological Process: iron ion homeostasis	Molecular Function: ferric iron binding	Molecular Function: 6-phosphofructokinase activity	Cellular Component: 6-phosphofructokinase complex	Cellular Component: cytoplasm	Molecular Function: microtubule motor activity	Cellular Component: microtubule associated complex	Biological Process: microtubule-based process
Q27778	6-Phosphofructokinase (Sm)	2	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	1	0	0	0
AAC14467	Cu,Zn-superoxide dismutase (Sm)	1b	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AAQ16180	Cysteine protease inhibitor (Sm)	1c	1	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0
AAL15461	Fatty acid-binding protein Sm14 (Sm)	1a	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
P25319	Ferritin-2 heavy chain (Sm)	3	0	0	0	0	1	0	0	0	0	0	1	1	1	0	0	0	0	0	0
Q94748	Probable dynein light chain (SM10) (T-cell-stimulating antigen SM10)	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1

Figure 28. GO terms corresponding to secretion-specific proteins



accession	description	phylo	IPR000120 Amidase	IPR012968 FerI	IPR008973 C2 calcium/lipid-binding region, CaLB	IPR012269 Aquaporin	IPR000425 Major intrinsic protein
TC10637	aquaporin	1a	0	0	0	1	1
CD157335	dysferlin	1b	0	1	1	0	0
CD156396	amidase	1c	1	0	0	0	0

accession	description	phylo	GO:0004040 Molecular Function: amidase activity	GO:0005215 Molecular Function: transporter activity	GO:0006810 Biological Process: transport	GO:0016021 Cellular Component: integral to membrane	GO:0016020 Cellular Component: membrane
TC10637	aquaporin	1a	0	1	1	1	1
CD156396	amidase	1c	1	0	0	0	0

Figure 29. Protein domains and GO terms corresponding to tegument-specific proteins

Figure 30. Phylogenetic tree of vesicle proteins

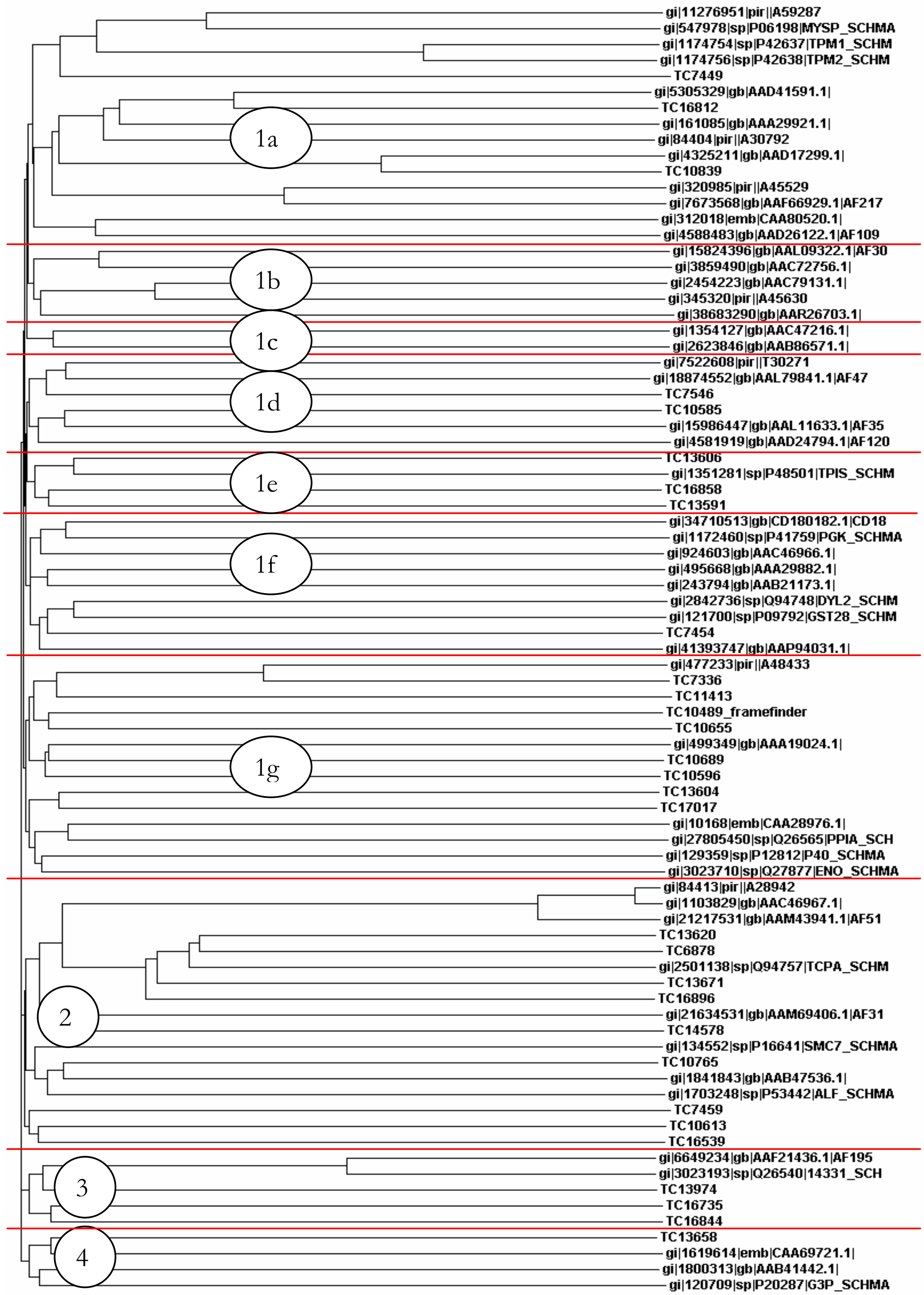

gi|11276951|pir||A59287
gi|547978|sp|P06198|MYSP_SCHMA
gi|1174754|sp|P42637|TPM1_SCHM
gi|1174756|sp|P42638|TPM2_SCHM
TC7449
gi|5305329|gb|AAD41591.1|
TC16812
gi|161085|gb|AAA29921.1|
1a
gi|84404|pir||A30792
gi|4325211|gb|AAD17299.1|
TC10839
gi|320985|pir||A45529
gi|7673568|gb|AAF66929.1|AF217
gi|312018|emb|CAA80520.1|
gi|4588483|gb|AAD26122.1|AF109
gi|15824396|gb|AAL09322.1|AF30
gi|3859490|gb|AAC72756.1|
1b
gi|2454223|gb|AAC79131.1|
gi|345320|pir||A45630
gi|38683290|gb|AAR26703.1|
1c
gi|1354127|gb|AAC47216.1|
gi|2623846|gb|AAB86571.1|
gi|7522608|pir||T30271
gi|18874552|gb|AAL79841.1|AF47
1d
TC7546
TC10585
gi|15986447|gb|AAL11633.1|AF35
gi|4581919|gb|AAD24794.1|AF120
TC13606
gi|1351281|sp|P48501|TPIS_SCHM
1e
TC16858
TC13591
gi|34710513|gb|CD180182.1|CD18
gi|1172460|sp|P41759|PGK_SCHMA
1f
gi|924603|gb|AAC46966.1|
gi|495668|gb|AAA29882.1|
gi|243794|gb|AAB21173.1|
gi|2842736|sp|Q94748|DYL2_SCHM
gi|121700|sp|P09792|GST28_SCHM
TC7454
gi|41393747|gb|AAP94031.1|
gi|477233|pir||A48433
TC7336
TC11413
TC10489_framefinder
TC10655
gi|499349|gb|AAA19024.1|
1g
TC10689
TC10596
TC13604
TC17017
gi|10168|emb|CAA28976.1|
gi|27805450|sp|Q26565|PPIA_SCH
gi|129359|sp|P12812|P40_SCHMA
gi|3023710|sp|Q27877|ENO_SCHMA
gi|84413|pir||A28942
gi|1103829|gb|AAC46967.1|
gi|21217531|gb|AAM43941.1|AF51
TC13620
TC6878
gi|2501138|sp|Q94757|TCPA_SCHM
TC13671
TC16896
2
gi|21634531|gb|AAM69406.1|AF31
TC14578
gi|134552|sp|P16641|SMC7_SCHMA
TC10765
gi|1841843|gb|AAB47536.1|
gi|1703248|sp|P53442|ALF_SCHMA
TC7459
TC10613
TC16539
gi|6649234|gb|AAF21436.1|AF195
gi|3023193|sp|Q26540|14331_SCH
3
TC13974
TC16735
TC16844
TC13658
4
gi|1619614|emb|CAA69721.1|
gi|1800313|gb|AAB41442.1|
gi|120709|sp|P20287|G3P_SCHMA

Figure 31. Phylogenetic tree of secretion proteins

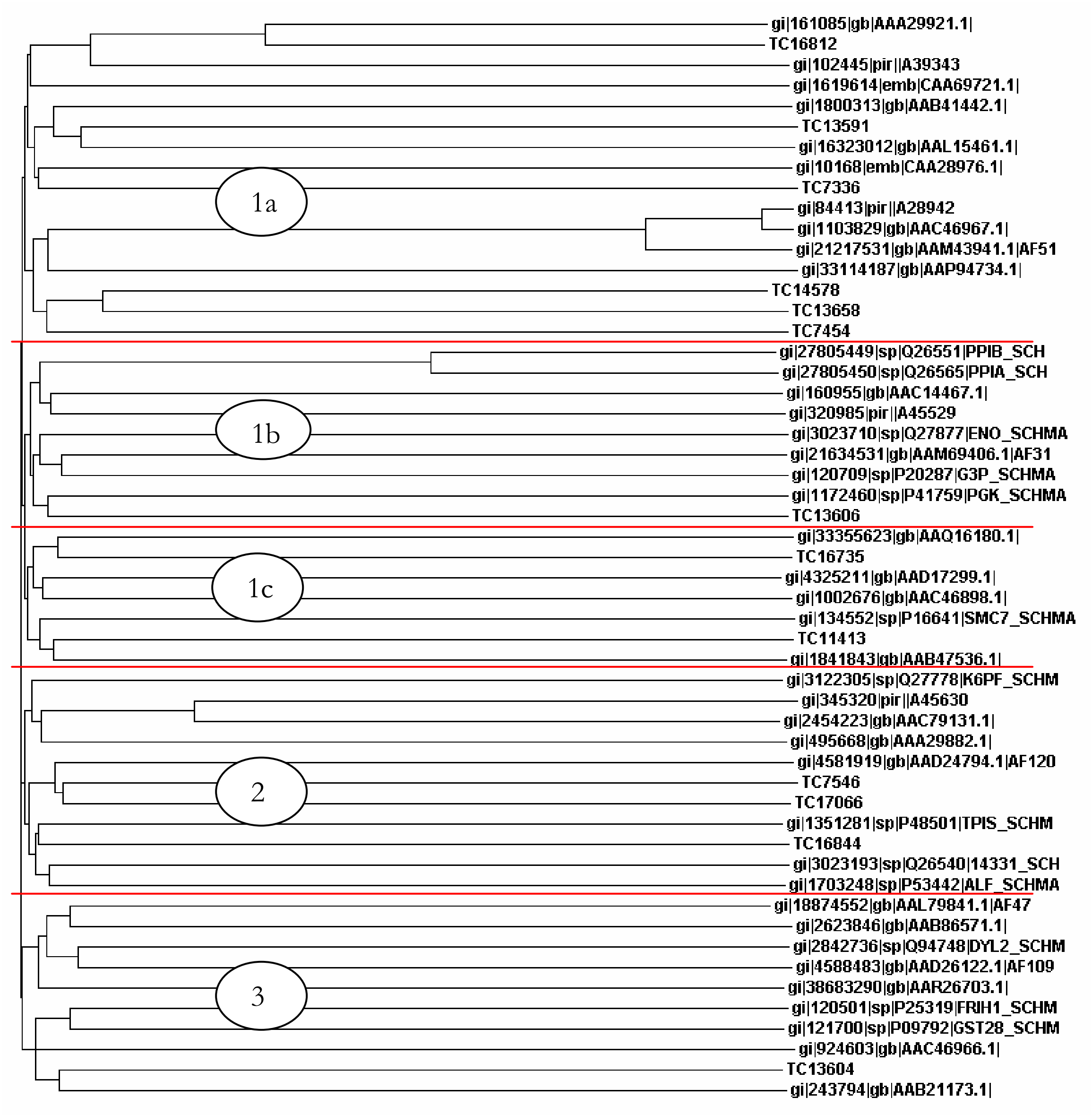

gi|161085|gb|AAA29921.1|
TC16812
gi|102445|pir||A39343
gi|1619614|emb|CAA69721.1|
gi|1800313|gb|AAB41442.1|
TC13591
gi|16323012|gb|AAL15461.1|
gi|10168|emb|CAA28976.1|
TC7336
1a
gi|84413|pir||A28942
gi|1103829|gb|AAC46967.1|
gi|21217531|gb|AAM43941.1|AF51
gi|33114187|gb|AAP94734.1|
TC14578
TC13658
TC7454
gi|27805449|sp|Q26551|PPIB_SCH
gi|27805450|sp|Q26565|PPIA_SCH
gi|160955|gb|AAC14467.1|
gi|320985|pir||A45529
1b
gi|3023710|sp|Q27877|ENO_SCHMA
gi|21634531|gb|AAM69406.1|AF31
gi|120709|sp|P20287|G3P_SCHMA
gi|1172460|sp|P41759|PGK_SCHMA
TC13606
gi|33355623|gb|AAQ16180.1|
TC16735
1c
gi|4325211|gb|AAD17299.1|
gi|1002676|gb|AAC46898.1|
gi|134552|sp|P16641|SMC7_SCHMA
TC11413
gi|1841843|gb|AAB47536.1|
gi|3122305|sp|Q27778|K6PF_SCHM
gi|345320|pir||A45630
gi|2454223|gb|AAC79131.1|
gi|495668|gb|AAA29882.1|
gi|4581919|gb|AAD24794.1|AF120
2
TC7546
TC17066
gi|1351281|sp|P48501|TPIS_SCHM
TC16844
gi|3023193|sp|Q26540|14331_SCH
gi|1703248|sp|P53442|ALF_SCHMA
gi|18874552|gb|AAL79841.1|AF47
gi|2623846|gb|AAB86571.1|
gi|2842736|sp|Q94748|DYL2_SCHM
gi|4588483|gb|AAD26122.1|AF109
3
gi|38683290|gb|AAR26703.1|
gi|120501|sp|P25319|FRIH1_SCHM
gi|121700|sp|P09792|GST28_SCHM
gi|924603|gb|AAC46966.1|
TC13604
gi|243794|gb|AAB21173.1|

Figure 32. Phylogenetic tree of tegument proteins

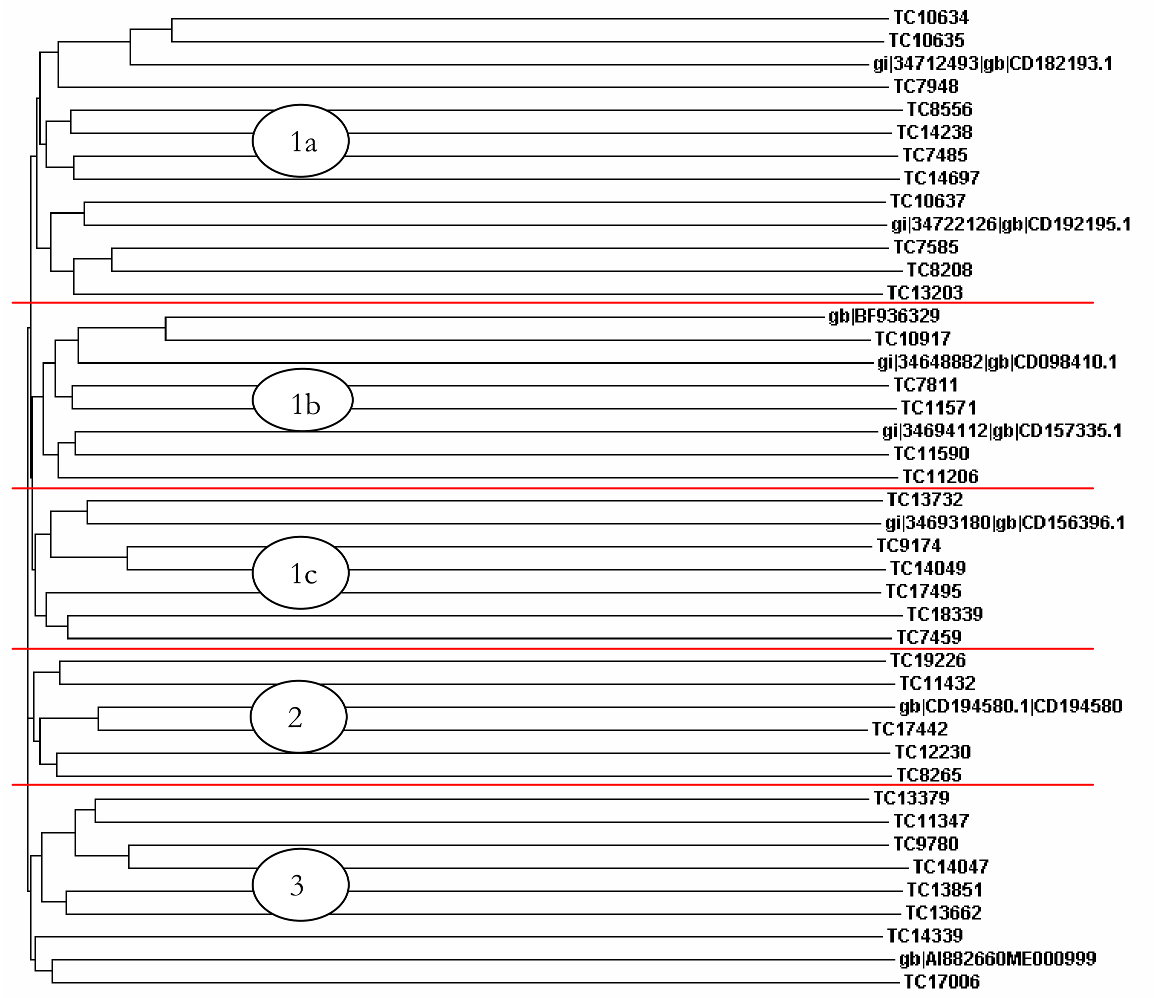


Table 8. GO terms and protein domains associated with phylogenetic groups

Sample Type	Phylo-genetic Group	Associated GO terms	Associated Protein Domains
Vesicles	1a	**Molecular Functions:** ATP binding, unfolded protein binding, calcium ion binding, motor activity, electron transporter activity, glutamate-ammonia ligase activity; **Biological Process:** protein folding, electron transport, ciliary or flagellar motility, nitrogen compound metabolism; **Cellular Component**: myosin, flagellum	Troponin; Calcium-binding EF-hand; Flagellar calcium-binding protein calflagin; Recoverin; EF-Hand type; Thioredoxin-like fold; Thioredoxin fold; Thioredoxin-related; Alkyl hydroperoxide reductase/ Thiol specific antioxidant/ Mal allergen; Tropomyosin; Myosin tail; Myosin, N-terminal, SH3-like; Myosin head, motor region; Endoplasmic reticulum targeting sequence; Four-helical cytokine; Heat shock protein Hsp90; Protein of unknown function DUF465; Glutamine synthetase class-I, adenylation site

Vesicles	1b	**Molecular Function:** catalytic activity, ATPase activity;
		Biological Process: metabolism, cation transport;
			Calcium-binding EF-hand; EF-Hand type; Filamin/ABP280 repeat; Carbohydrate kinase, PfkB; Haloacid dehalogenase-like hydrolase; Cation transporting ATPase, N-terminal; Cation transporting ATPase, C-terminal; ATPase, E1-E2 type; E1-E2 ATPase-associated region; Cation transporting ATPase;

Vesicles	1c	**Molecular Function:** serine-type endopeptidase inhibitor activity; **Biological Process:** cell adhesion;	Beta-Ig-H3/fasciclin; Proteinase inhibitor I4, serpin;

Sample Type	Phylo-genetic Group	Associated GO terms	Associated Protein Domains
Vesicles	1d	**Molecular Function:** intramolecular transferase activity, ATP binding, GTP binding, hydrogen-transporting ATP synthase activity, hydrogen-transporting ATPase activity, hydrolase activity, electron transporter activity, molecular function unknown, phosphoenolpyruvate carboxykinase activity; **Biological Process:** glycolysis, electron transport, ATP synthesis coupled proton transport, ATP biosynthesis, gluconeogenesis; **Cellular Component**: cytoplasm, proton-transporting two-sector ATPase complex;	Phosphoglycerate mutase; Phosphoglycerate mutase 1 H+-transporting two-sector ATPase, alpha/beta subunit, central region; H+-transporting two-sector ATPase, alpha/beta subunit, N-terminal; H+-transporting two-sector ATPase, alpha/beta subunit, C-terminal; Thioredoxin fold; ATP synthase F1, alpha subunit; Thioredoxin-related; Translationally controlled tumor protein; Mss4-like; Phosphoenolpyruvate carboxykinase GTP;
Vesicles	1e	**Molecular Function:** binding, transporter activity, DNA binding, phosphorylase activity, pyridoxal phosphate binding, triose-phosphate isomerase activity; **Biological Process:** transport, nucleosome assembly, chromosome organization and biogenesis [sensu Eukaryota], carbohydrate metabolism, metabolism **Cellular Component:** membrane, mitochondrial inner membrane, nucleosome, nucleus;	Mitochondrial substrate carrier; Adenine nucleotide translocator 1; Mitochondrial carrier protein; Histone core; Histone-fold; Histone H2B; Glycosyl transferase, family 35; Glycogen/starch/alpha-glucan phosphorylase; Triosephosphate isomerase;
Vesicles	1f	**Molecular Function:** ATP binding, pyruvate kinase activity, protein binding, microtubule motor activity, phosphoglycerate kinase activity, motor activity, protein kinase activity, arginase activity, catalytic activity, structural constituent of cytoskeleton, actin binding; **Biological Process:** glycolysis, microtubule-based process, protein amino acid phosphorylation, arginine catabolism; **Cellular Component:** microtubule associated complex, actin filament;	Pyruvate kinase; Thioredoxin fold; Dynein light chain, type 1; Phosphoglycerate kinase; Glutathione S-transferase, C-terminal-like; Glutathione S-transferase, C-terminal; Glutathione S-transferase, N-terminal; Protein kinase; Protein kinase-like; Arginase; Arginase/agmatinase/formiminoglutamase; Actin; Actin/actin-like; Calponin-like actin-binding; Actin-binding, actinin-type;

Sample Type	Phylo-genetic Group	Associated GO terms	Associated Protein Domains
Vesicles	1g	**Molecular Function:** nucleic acid binding, ATP binding, GTP binding, GTPase activity, structural molecule activity, unfolded protein binding, enzyme inhibitor activity, calcium ion binding, hydrogen-transporting ATP synthase activity, hydrogen-transporting ATPase activity, nucleotide binding, nucleoside-triphosphatase activity, hydrogen-exporting ATPase activity, nucleoside diphosphate kinase activity, magnesium ion binding, transferase activity, citrate [Si]-synthase activity, hydrolase activity, phosphopyruvate hydratase activity, **Biological Process:** glycolysis, protein polymerization, microtubule-based movement, protein folding, negative regulation of nucleoside metabolism, ATP synthesis coupled proton transport, ATP biosynthesis, GTP biosynthesis, UTP biosynthesis, CTP biosynthesis, pyrimidine ribonucleoside triphosphate biosynthesis, main pathways of carbohydrate metabolism, tricarboxylic acid cycle; **Cellular Component:** cytoplasm, protein complex, microtubule, mitochondrion, proton-transporting two-sector ATPase complex, integral to membrane, hydrogen-translocating F-type ATPase complex, phosphopyruvate hydratase complex, endoplasmic reticulum	Tubulin/FtsZ, C-terminal; Tubulin; Tubulin/FtsZ, GTPase; Beta tubulin; Mitochondrial ATPase inhibitor, IATP; H+-transporting two-sector ATPase, alpha/beta subunit, central region; AAA ATPase; H+-transporting two-sector ATPase, alpha/beta subunit, N-terminal; H+-transporting two-sector ATPase, alpha/beta subunit, C-terminal; ATP synthase F1, beta subunit; Nucleoside diphosphate kinase; Nucleoside-diphosphate kinase; Band 7 protein; Prohibitin; Citrate synthase; Citrate synthase, eukaryotic; AAA-protein subdomain; AAA ATPase VAT, N-terminal; Aspartate decarboxylase-like fold; AAA ATPase, central region; AAA ATPase, CDC48; RNA-binding region RNP-1 RNA recognition motif; Nucleotide-binding, alpha-beta plait; Enolase; Peptidyl-prolyl cis-trans isomerase, cyclophilin type;HSP20-like chaperone; Alpha crystallin; Heat shock protein Hsp20; Heat shock protein 70; Heat shock protein Hsp70; Calreticulin/calnexin; Concanavalin A-like lectin/glucanase, subgroup; Concanavalin A-like lectin/glucanase; Calreticulin; Alpha tubulin

Sample Type	Phylo-genetic Group	Associated GO terms	Associated Protein Domains
Vesicles	2	**Molecular Function:** intramolecular transferase activity; nucleic acid binding; helicase activity; ATP binding; protein binding; structural constituent of ribosome; fructose-bisphosphate aldolase activity; kinase activity; transferase activity; serine-type peptidase activity; serine-type endopeptidase activity; actin binding; unfolded protein binding; transketolase activity; DNA binding; **Biological Process:** glycolysis; protein folding; anaerobic glycolysis; cellular protein metabolism; nucleosome assembly; chromosome organization and biogenesis [sensu Eukaryota]; carbohydrate metabolism; protein biosynthesis; proteolysis; actomyosin structure organization and biogenesis **Cellular Component:** nucleosome; nucleus; intracellular; ribosome;	Helicase, C-terminal; DEAD/DEAH box helicase, N-terminal; Chaperonin TCP-1; Chaperonin Cpn60; Chaperonin Cpn60/TCP-1; GroEL-like chaperone, ATPase; T-complex protein 1, epsilon subunit; T-complex protein 1, zeta subunit; L-lactate dehydrogenase; Transketolase, central region; Transketolase, C-terminal-like; Transketolase, C-terminal; Transketolase, N-terminal; Bacterial transketolase; Histone core; Histone-fold; Histone H4; T-complex protein 1, gamma subunit; T-complex protein 1, beta subunit; Ribosomal protein L11; Phosphoglucomutase/phosphomannomutase; Phosphoglucomutase/phosphomannomutase C terminal; Phosphoglucomutase/phosphomannomutase alpha/beta/alpha domain II; Phosphoglucomutase/phosphomannomutase alpha/beta/alpha domain III; Phosphoglucomutase/phosphomannomutase alpha/beta/alpha domain I; Fructose-bisphosphate aldolase, class-I; ATP:guanido phosphotransferase; Peptidase S1B, glutamyl endopeptidase I; Peptidase S1A, chymotrypsin; Peptidase S1 and S6, chymotrypsin/Hap; Peptidase, trypsin-like serine and cysteine; Calponin repeat; Calponin; Calponin-like actin-binding; SM22/calponin
Vesicles	3	**Molecular Function:** malate dehydrogenase activity, L-lactate dehydrogenase activity, oxidoreductase activity, L-malate dehydrogenase activity, protein domain specific binding **Biological Process:** glycolysis, malate metabolism, tricarboxylic acid cycle intermediate metabolism, anaerobic glycolysis, electron transport; **Cellular Component:** cytoplasm;	Malate dehydrogenase, active site; L-lactate/malate dehydrogenase; Lactate/malate dehydrogenase; Malate dehydrogenase, NAD-dependent, eukaryotes and gamma proteobacteria; L-lactate dehydrogenase; Fumarate reductase/succinate dehydrogenase flavoprotein, N-terminal; Fumarate reductase/succinate dehydrogenase flavoprotein, C-terminal; 14-3-3 protein;

Sample Type	Phylo-genetic Group	Associated GO terms	Associated Protein Domains
Vesicles	4	**Molecular Function:** GTP binding, DNA binding, glyceraldehyde-3-phosphate dehydrogenase [phosphorylating] activity, NAD binding, aminopeptidase activity, leucyl aminopeptidase activity, manganese ion binding **Biological Process:** glycolysis, nucleosome assembly, chromosome organization and biogenesis [sensu Eukaryota], protein biosynthesis, proteolysis, protein metabolism; **Cellular Component:** cytoplasm, nucleosome, nucleus, intracellular;	Histone core; Histone-fold; Histone H3; Glyceraldehyde 3-phosphate dehydrogenase; Elongation factor Tu, domain 2; Translation factor; Protein synthesis factor, GTP-binding; EF-Tu/eEF-1alpha/eIF2-gamma, C-terminal; Elongation factor Tu, C-terminal; Peptidase M17, leucyl aminopeptidase, C-terminal; Peptidase M17, leucyl aminopeptidase
Secretions	1a	**Molecular Function:** serine-type endopeptidase activity, calcium ion binding, aminopeptidase activity, GTP binding, binding, lipid binding, serine-type peptidase activity, ATP binding, hydrolase activity, phosphorylase activit, pyridoxal phosphate binding, DNA binding, structural molecule activity, pyruvate kinase activity **Biological Process:** proteolysis, ciliary or flagellar motility, transport, protein biosynthesis, glycolysis, carbohydrate metabolism, nucleosome assembly, chromosome organization and biogenesis, microtubule-based movement, GTPase activity, protein polymerization **Cellular Component:** intracellular, flagellum, cytoplasm, nucleosome, nucleus, microtubule, protein complex	Peptidase S1 and S6, chymotrypsin/Hap, Peptidase, trypsin-like serine and cysteine, Peptidase S1A, chymotrypsin, Peptidase C2, calpain, EF-Hand type, Peptidase, cysteine peptidase active site, Calcium-binding EF-hand, Flagellar calcium-binding protein calflagin, Peptidase M17, leucyl aminopeptidase, C-terminal, Cytosolic fatty-acid binding, Calycin, Lipocalin-related protein and Bos/Can/Equ allergen, Calycin-like, Peptidase S1B, glutamyl endopeptidase I, Nucleoside phosphatase GDA1/CD39, Heat shock protein Hsp70, Heat shock protein 70, Protein synthesis factor, GTP-binding, Elongation factor Tu, C-terminal, EF-Tu/eEF-1alpha/eIF2-gamma, C-terminal, Translation factor, Elongation factor Tu, domain 2, Glycosyl transferase, family 35, Glycogen/starch/alpha-glucan phosphorylase, Histone-fold, Histone core, Histone H3, Histone H4, Recoverin, Tubulin, Beta tubulin, Tubulin/FtsZ, GTPase, Tubulin/FtsZ, C-terminal, Cell division protein FtsZ, Pyruvate kinase

Sample Type	Phylo-genetic Group	Associated GO terms	Associated Protein Domains
Secretions	1b	**Molecular Function**: unfolded protein binding, copper, zinc superoxide dismutase activity, metal ion binding, protein binding, ATP binding, glyceraldehyde-3-phosphate dehydrogenase, phosphorylating activity, NAD binding, phosphoglycerate kinase activity, phosphopyruvate hydratase activity, DNA binding **Biological Process**: protein folding, superoxide metabolism, cellular protein metabolism, glycolysis, nucleosome assembly, chromosome organization and biogenesis **Cellular Component**: phosphopyruvate hydratase complex, nucleosome, nucleus	Heat shock protein Hsp90, Four-helical cytokine, Superoxide dismutase, copper/zinc binding, Chaperonin Cpn60, Chaperonin Cpn60/TCP-1, GroEL-like chaperone, ATPase, Glyceraldehyde 3-phosphate dehydrogenase, Phosphoglycerate kinase, Peptidyl-prolyl cis-trans isomerase, cyclophilin type, Enolase, Histone H2B, Histone-fold, Histone core,
Secretions	1c	**Molecular Function:** calcium ion binding, oxidoreductase activity, ATP binding, cysteine protease inhibitor activity, endopeptidase inhibitor activity, kinase activity, transferase activity, nucleoside diphosphate kinase activity, magnesium ion binding, L-lactate dehydrogenase activity **Biological Process:** actomyosin structure organization and biogenesis, metabolism, glycolysis, GTP biosynthesis, UTP biosynthesis, CTP biosynthesis, pyrimidine ribonucleoside triphosphate biosynthesis, anaerobic glycolysis, tricarboxylic acid cycle intermediate metabolism **Cellular Component:** intracellular, cytoplasm	Calponin-like actin-binding, Thioredoxin fold, Calponin repeat, SM22/calponin, Calponin, Glucose/ribitol dehydrogenase, Alkyl hydroperoxide reductase/ Thiol specific antioxidant/ Mal allergen, Proteinase inhibitor I25, cystatin, Proteinase inhibitor I25A, stefin A,ATP:guanido phosphotransferase, Nucleoside diphosphate kinase, Nucleoside-diphosphate kinase, L-lactate dehydrogenase, Lactate/malate dehydrogenase, L-lactate/malate dehydrogenase

Sample Type	Phylo-genetic Group	Associated GO terms	Associated Protein Domains
Secretions	2	**Molecular Function**: calcium ion binding, calcium ion binding, oxidoreductase activity, phosphoenolpyruvate carboxykinase activity, GTP binding **Biological Process**: metabolism, gluconeogenesis, glycolysis, triose-phosphate isomerase activity, fructose-bisphosphate aldolase activity, protein domain specific binding, 6-phosphofructokinase activity **Cellular Component**: 6-phosphofructokinase complex, cytoplasm, L-lactate dehydrogenase activity, tricarboxylic acid cycle intermediate metabolism, malate metabolism, L-malate dehydrogenase activity, malate dehydrogenase activity, intramolecular transferase activity	EF-Hand type, Calcium-binding EF-hand, Actin-binding, actinin-type, Calponin-like actin-binding, Phosphoenolpyruvate carboxykinase GTP, Triosephosphate isomerase, Fructose-bisphosphate aldolase, class-I, 14-3-3 protein, Phosphofructokinase, 6-phosphofructokinase, eukaryotic type, Lactate/malate dehydrogenase, L-lactate/malate dehydrogenase, Malate dehydrogenase, NAD-dependent, eukaryotes and gamma proteobacteria, Malate dehydrogenase, active site, Malate dehydrogenase, NAD-dependent, cytosolic, Malate dehydrogenase, NAD or NADP, Malate dehydrogenase, Phosphoglycerate mutase 1, Phosphoglycerate mutase
Secretions	3	**Molecular Function:** serine-type endopeptidase inhibitor activity, protein binding, motor activity, structural constituent of cytoskeleton, binding, electron transporter activity, ATP binding, ferric iron binding, microtubule motor activity, hydrogen-exporting ATPase activity, hydrogen-transporting ATP synthase activity, hydrogen-transporting ATPase activity, nucleotide binding, nucleoside-triphosphatase activity **Biological Process:** ATP synthesis coupled proton transport, electron transport, iron ion transport, iron ion homeostasis, microtubule-based process, ATP biosynthesis **Cellular Component:** actin filament, microtubule associated complex, integral to membrane, hydrogen-translocating F-type ATPase complex, proton-transporting two-sector ATPase complex	Glutathione S-transferase, C-terminal, Glutathione S-transferase, C-terminal-like, Glutathione S-transferase, N-terminal, Thioredoxin fold, Proteinase inhibitor I4, serpin, Actin/actin-like, Actin, Thioredoxin-related, Carbohydrate kinase, PfkB, Filamin/ABP280 repeat, Ferritin, Ferritin-related, Ferritin and Dps, Ferritin/ribonucleotide reductase-like, Dynein light chain, type 1, ATP synthase F1, beta subunit, H+-transporting two-sector ATPase, alpha/beta subunit, central region, H+-transporting two-sector ATPase, alpha/beta subunit, C-terminal, H+-transporting two-sector ATPase, alpha/beta subunit, N-terminal, AAA ATPase

Sample Type	Phylo-genetic Group	Associated GO terms	Associated Protein Domains
Tegument	1a	**Molecular Function:** transporter activity, protein binding **Biological Process:** transport **Cellular Component:** integral to membrane, membrane, extracellular region	C2 calcium/lipid-binding region, CaLB, Allergen V5/Tpx-1 related, Aquaporin, Major intrinsic protein, CD9/CD37/CD63 antigen, Tetraspanin, PDZ/DHR/GLGF, Calponin-like actin-binding, HSP20-like chaperone, Heat shock protein Hsp20, Alpha crystallin,
Tegument	1b	**Molecular Function:** calcium ion binding, calcium-dependent phospholipid binding, protein-L-isoaspartate [D-aspartate] O-methyltransferase activity, structural constituent of ribosome **Biological Process:** protein modification, protein biosynthesis, protein targeting **Cellular Component:** intracellular, ribosome, Golgi stack	Annexin, FerI, C2 calcium/lipid-binding region, CaLB, Protein-L-isoaspartateD-aspartate O-methyltransferase, Thioredoxin-like fold, Ubiquitin, Ribosomal protein L40e, Protein of unknown function DUF1606, AP2 clathrin adaptor, alpha and beta chain, appendage
Tegument	1c	**Molecular Function:** amidase activity, catalytic activity, nucleic acid binding, helicase activity, ATP binding **Biological Process:** metabolism **Cellular Component:** integral to membrane	Amidase, Thioredoxin-like fold, CD9/CD37/CD63 antigen, Tetraspanin, Alkyl hydroperoxide reductase/ Thiol specific antioxidant/ Mal allergen, Thioredoxin fold, AMP-dependent synthetase and ligase, DEAD/DEAH box helicase, N-terminal, Helicase, C-terminal
Tegument	2	**Molecular Function:** transporter activity, hydrolase activity, sugar porter activity **Biological Process:** transport, carbohydrate transport **Cellular Component:** integral to membrane, membrane, intracellular	Nucleoside phosphatase GDA1/CD39, Phosphatidylinositol transfer protein, General substrate transporter, Sugar transporter, Sugar transporter superfamily, CAP protein

Sample Type	Phylo-genetic Group	Associated GO terms	Associated Protein Domains
Tegument	3	**Molecular Function:** calcium ion binding, hydrolase activity, molecular function unknown **Biological Process:** exocytosis, vesicle docking, nucleotide metabolism **Cellular Component:** cytoplasm	Exocyst complex component Sec10, Snf7, Spectrin repeat, Calcium-binding EF-hand, EF-Hand type, Type I phosphodiesterase/nucleotide pyrophosphatase, Filamin/ABP280 repeat, Calponin-like actin-binding

Protein Domain/Gene Ontology Annotation and Statistical Analysis

All proteins in each of the sample sets were annotated with GO terms and protein domains using InterProScan. Domains and GO terms with an e-value less than 1×10^{-5} were considered significant. A partial example of the raw results generated by InterProScan is provided in Figure 33. A Perl script (Appendix J) was written to parse the raw results into 1) counts of GO terms (Appendix F), 2) counts of protein domains (Appendix G), 3) matrices of sequence identifier vs. GO terms, and 4) matrices of sequence identifier vs. protein domains. The matrices were used as input files for a Matlab script (Appendix J) that computed the entropy of each GO term (Appendix H) and protein domain (Appendix I).

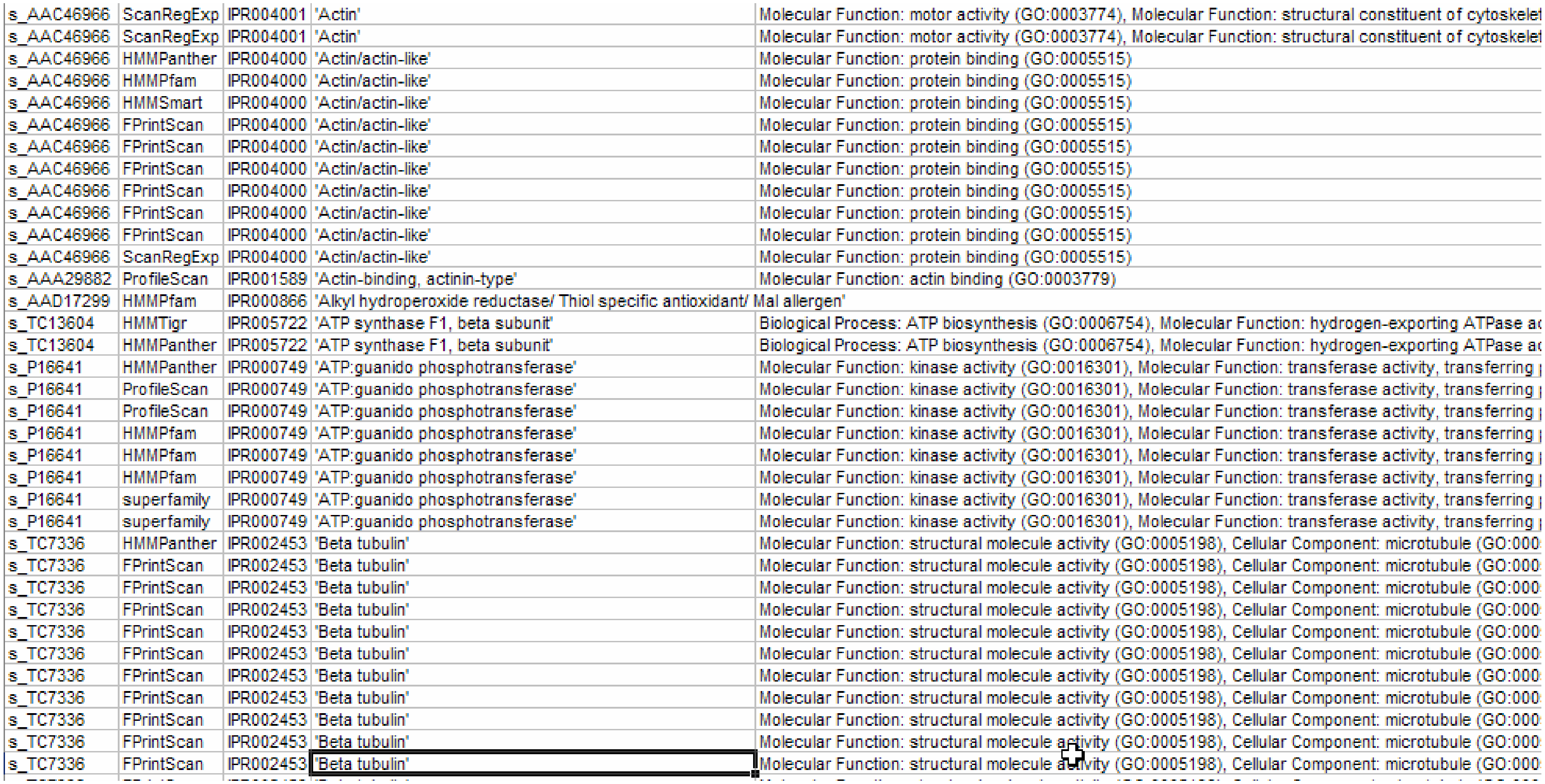
Entropy was calculated for each attribute (protein domain or GO id) and used to determine the most statistically significant GO terms and protein domains. The overall information (I) contained in a data set is equal to the amount of information needed to decide if an arbitrary sample in S belongs to P or N, for example. Entropy (E(A)) represents the expected information needed to classify objects in all subtrees, while information gain (G(A)) represents the encoding information that would be gained by branching on A. The formulas used in these calculations are as follows:

$$I(p,n) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n}$$

$$E(A) = \sum_{i=1}^{v}\frac{p_i+n_i}{p+n}I(p_i,n_i)$$

$$Gain(A) = I(p,n) - E(A)$$

Low entropy values correspond to high information gain values. In the entropy calculation tables in Appendix H, the columns “yy”, “ny”, “nn”, and “yn” indicate how many times a GO term or protein domain occurred in both samples, one but not the other, and neither sample. The table headings correspond to these entries; the sample type that is listed first



s_AAC46966	ScanRegExp	IPR004001	'Actin'	Molecular Function: motor activity (GO:0003774), Molecular Function: structural constituent of cytoskelet
s_AAC46966	ScanRegExp	IPR004001	'Actin'	Molecular Function: motor activity (GO:0003774), Molecular Function: structural constituent of cytoskelet
s_AAC46966	HMMPanther	IPR004000	'Actin/actin-like'	Molecular Function: protein binding (GO:0005515)
s_AAC46966	HMMPfam	IPR004000	'Actin/actin-like'	Molecular Function: protein binding (GO:0005515)
s_AAC46966	HMMSmart	IPR004000	'Actin/actin-like'	Molecular Function: protein binding (GO:0005515)
s_AAC46966	FPrintScan	IPR004000	'Actin/actin-like'	Molecular Function: protein binding (GO:0005515)
s_AAC46966	FPrintScan	IPR004000	'Actin/actin-like'	Molecular Function: protein binding (GO:0005515)
s_AAC46966	FPrintScan	IPR004000	'Actin/actin-like'	Molecular Function: protein binding (GO:0005515)
s_AAC46966	FPrintScan	IPR004000	'Actin/actin-like'	Molecular Function: protein binding (GO:0005515)
s_AAC46966	FPrintScan	IPR004000	'Actin/actin-like'	Molecular Function: protein binding (GO:0005515)
s_AAC46966	FPrintScan	IPR004000	'Actin/actin-like'	Molecular Function: protein binding (GO:0005515)
s_AAC46966	ScanRegExp	IPR004000	'Actin/actin-like'	Molecular Function: protein binding (GO:0005515)
s_AAA29882	ProfileScan	IPR001589	'Actin-binding, actinin-type'	Molecular Function: actin binding (GO:0003779)
s_AAD17299	HMMPfam	IPR000866	'Alkyl hydroperoxide reductase/ Thiol specific antioxidant/ Mal allergen'	
s_TC13604	HMMTigr	IPR005722	'ATP synthase F1, beta subunit'	Biological Process: ATP biosynthesis (GO:0006754), Molecular Function: hydrogen-exporting ATPase a
s_TC13604	HMMPanther	IPR005722	'ATP synthase F1, beta subunit'	Biological Process: ATP biosynthesis (GO:0006754), Molecular Function: hydrogen-exporting ATPase a
s_P16641	HMMPanther	IPR000749	'ATP:guanido phosphotransferase'	Molecular Function: kinase activity (GO:0016301), Molecular Function: transferase activity, transferring
s_P16641	ProfileScan	IPR000749	'ATP:guanido phosphotransferase'	Molecular Function: kinase activity (GO:0016301), Molecular Function: transferase activity, transferring
s_P16641	ProfileScan	IPR000749	'ATP:guanido phosphotransferase'	Molecular Function: kinase activity (GO:0016301), Molecular Function: transferase activity, transferring
s_P16641	HMMPfam	IPR000749	'ATP:guanido phosphotransferase'	Molecular Function: kinase activity (GO:0016301), Molecular Function: transferase activity, transferring
s_P16641	HMMPfam	IPR000749	'ATP:guanido phosphotransferase'	Molecular Function: kinase activity (GO:0016301), Molecular Function: transferase activity, transferring
s_P16641	HMMPfam	IPR000749	'ATP:guanido phosphotransferase'	Molecular Function: kinase activity (GO:0016301), Molecular Function: transferase activity, transferring
s_P16641	superfamily	IPR000749	'ATP:guanido phosphotransferase'	Molecular Function: kinase activity (GO:0016301), Molecular Function: transferase activity, transferring
s_P16641	superfamily	IPR000749	'ATP:guanido phosphotransferase'	Molecular Function: kinase activity (GO:0016301), Molecular Function: transferase activity, transferring
s_TC7336	HMMPanther	IPR002453	'Beta tubulin'	Molecular Function: structural molecule activity (GO:0005198), Cellular Component: microtubule (GO:000
s_TC7336	FPrintScan	IPR002453	'Beta tubulin'	Molecular Function: structural molecule activity (GO:0005198), Cellular Component: microtubule (GO:000
s_TC7336	FPrintScan	IPR002453	'Beta tubulin'	Molecular Function: structural molecule activity (GO:0005198), Cellular Component: microtubule (GO:000
s_TC7336	FPrintScan	IPR002453	'Beta tubulin'	Molecular Function: structural molecule activity (GO:0005198), Cellular Component: microtubule (GO:000
s_TC7336	FPrintScan	IPR002453	'Beta tubulin'	Molecular Function: structural molecule activity (GO:0005198), Cellular Component: microtubule (GO:000
s_TC7336	FPrintScan	IPR002453	'Beta tubulin'	Molecular Function: structural molecule activity (GO:0005198), Cellular Component: microtubule (GO:000
s_TC7336	FPrintScan	IPR002453	'Beta tubulin'	Molecular Function: structural molecule activity (GO:0005198), Cellular Component: microtubule (GO:000
s_TC7336	FPrintScan	IPR002453	'Beta tubulin'	Molecular Function: structural molecule activity (GO:0005198), Cellular Component: microtubule (GO:000
s_TC7336	FPrintScan	IPR002453	'Beta tubulin'	Molecular Function: structural molecule activity (GO:0005198), Cellular Component: microtubule (GO:000
s_TC7336	FPrintScan	IPR002453	'Beta tubulin'	Molecular Function: structural molecule activity (GO:0005198), Cellular Component: microtubule (GO:000
s_TC7336	FPrintScan	IPR002453	'Beta tubulin'	Molecular Function: structural molecule activity (GO:0005198), Cellular Component: microtubule (GO:000

Figure 33. InterProScan raw data file example

corresponds to the first value in each of those four columns (y, n, n, y). For example, if the table heading indicates the calculations are for “vesicles vs. secretions”, then the “ny” column corresponds to a GO term not occurring in the vesicles but occurring in the secretions.

Gene Ontology Analysis

Detailed GO annotation data are provided in Appendix F and visual representations of categories Biological Process, Molecular Function (I and II), and Cellular Component are provided in Figures 34-37 respectively.

The most represented GO terms in the vesicle proteins in the category Biological Process include protein folding (12.35%), glycolysis (9.88%), cellular protein metabolism (7.41%) and proteolysis (4.94%). The most represented term in the Cellular Component category is cytoplasm (6.17%). The most represented Molecular Function is by far ATP binding at 19.75%, followed by unfolded protein binding (11.11%), calcium ion binding (8.64%) and protein binding (8.64%).

The secretions proteins have glycolysis as the most represented term under Biological Process (18.87%) followed by proteolysis (9.43%) and protein folding (7.55%). Under the Cellular Component category intracellular, cytoplasm, nucleosome, and nucleus are all evenly represented at 5.66%. Under the Molecular Function category calcium ion binding is most represented (9.43%) followed by oxidoreductase activity and ATP binding (both 7.55%). Secretions have a slightly higher representation of catalytic and enzymatic activity than vesicles (5.6% vs. 3.7%).

Not surprisingly, the tegument proteins have transport as the most represented Biological Process (6.98%) followed by protein modification (4.65%). Under Cellular Component integral to membrane (9.3%) and membrane (6.98%) are most represented. Under Molecular Function calcium ion binding, transporter activity, and hydrolase activity are equally represented (4.65%).

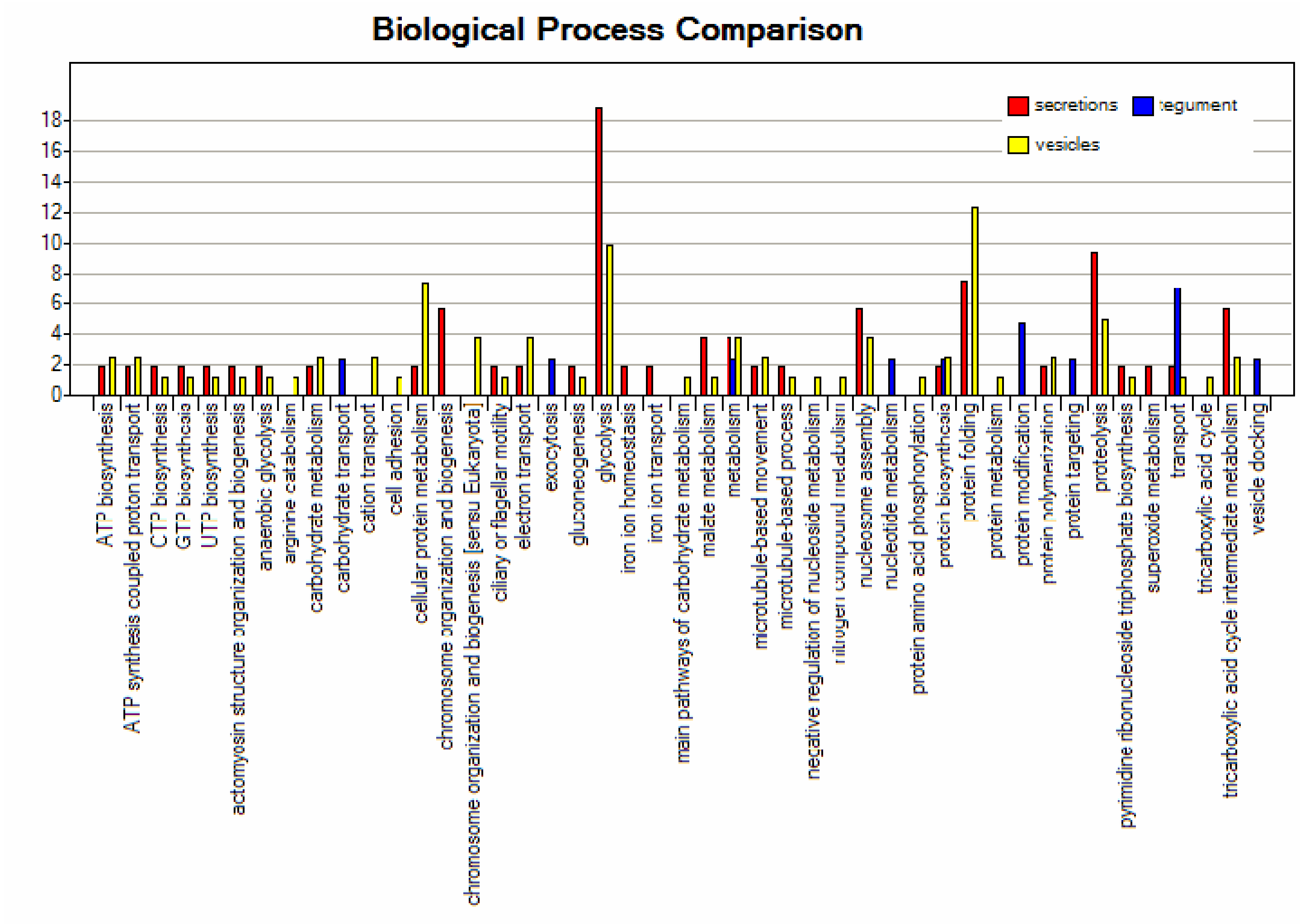


Figure 34. % of sequences corresponding to GO annotations from category biological process

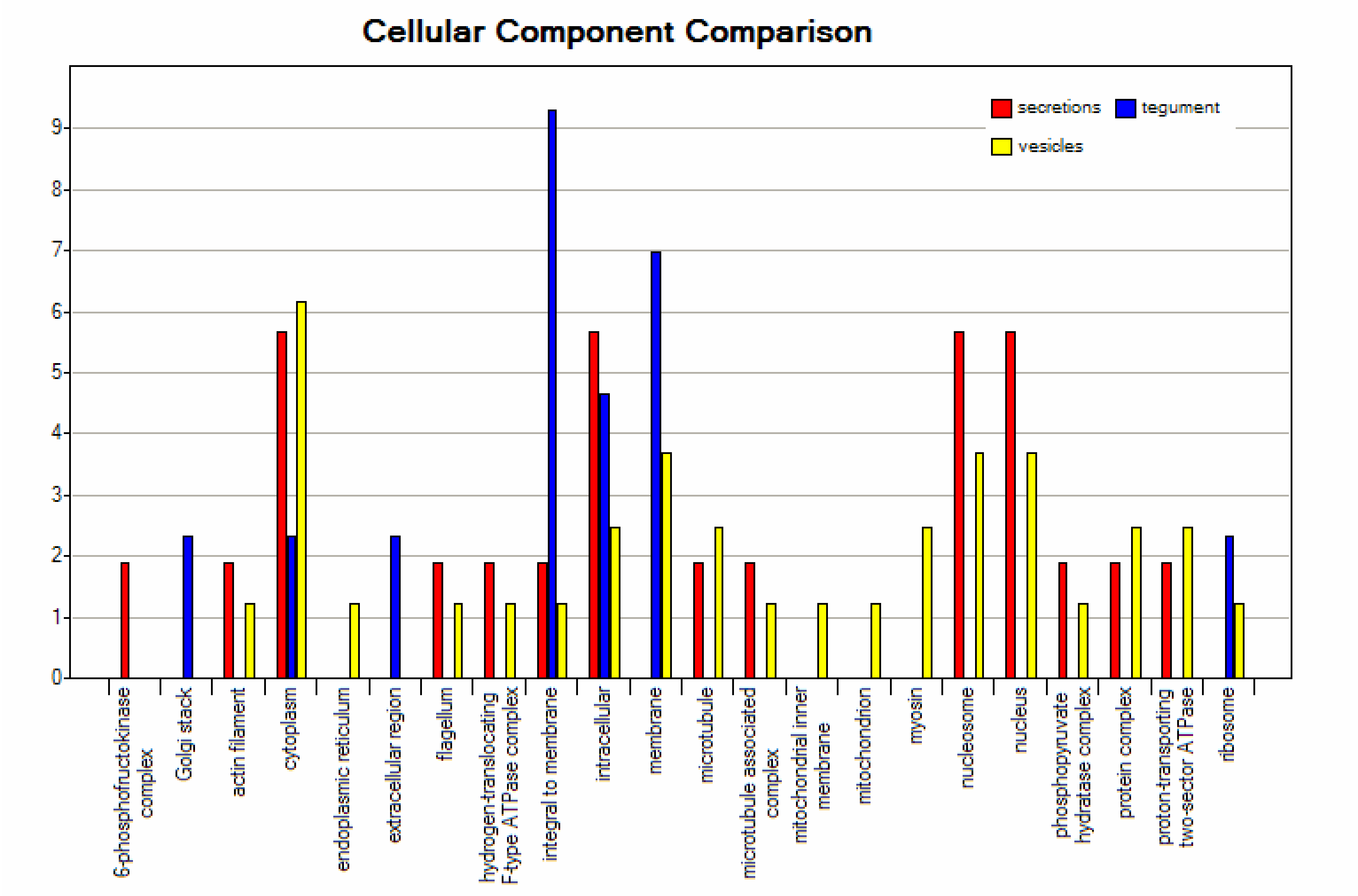


Figure 35. % of sequences corresponding to GO annotations from category cellular component

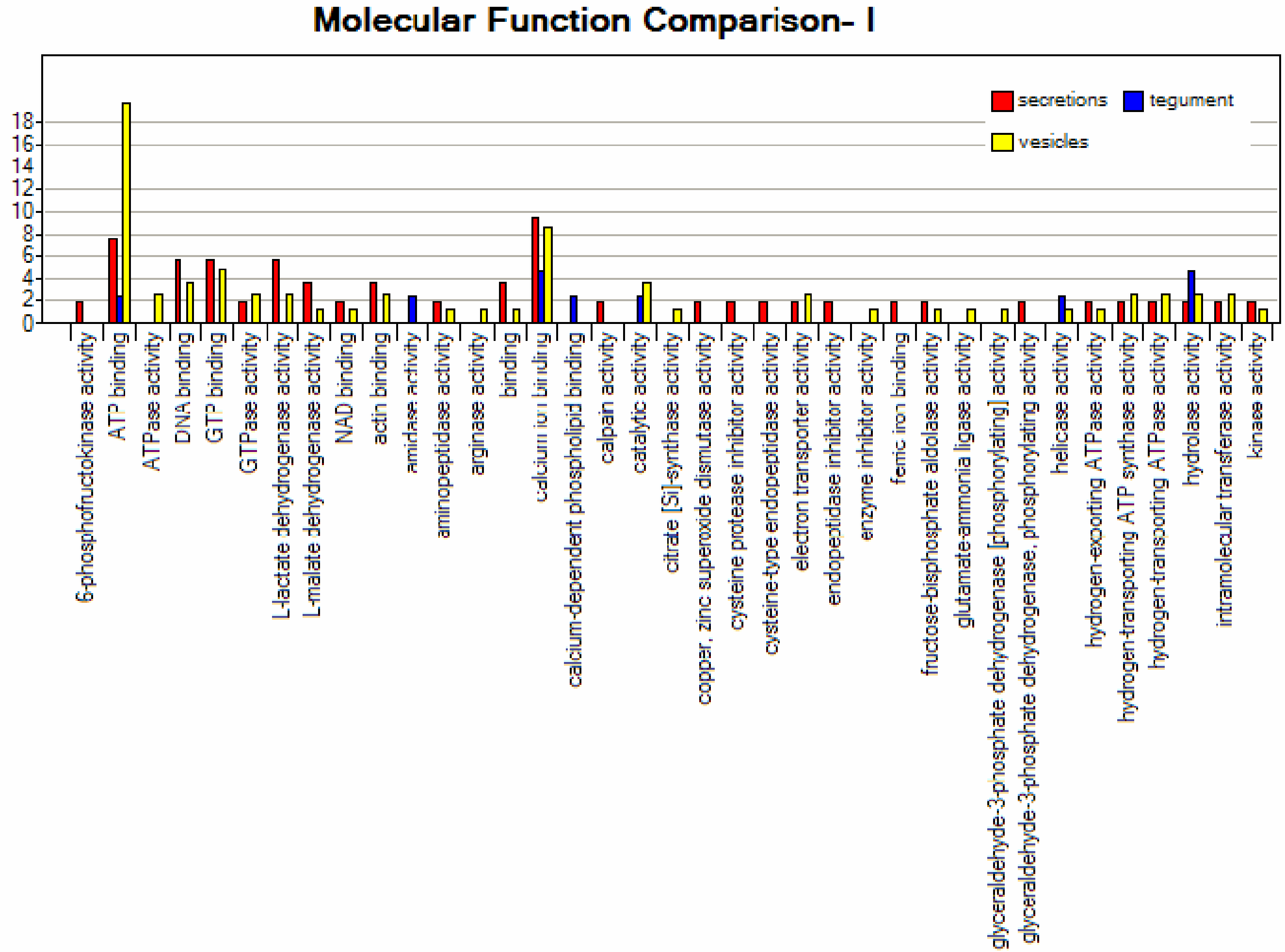


Figure 36. % of sequences corresponding to GO annotations from category molecular function (I)

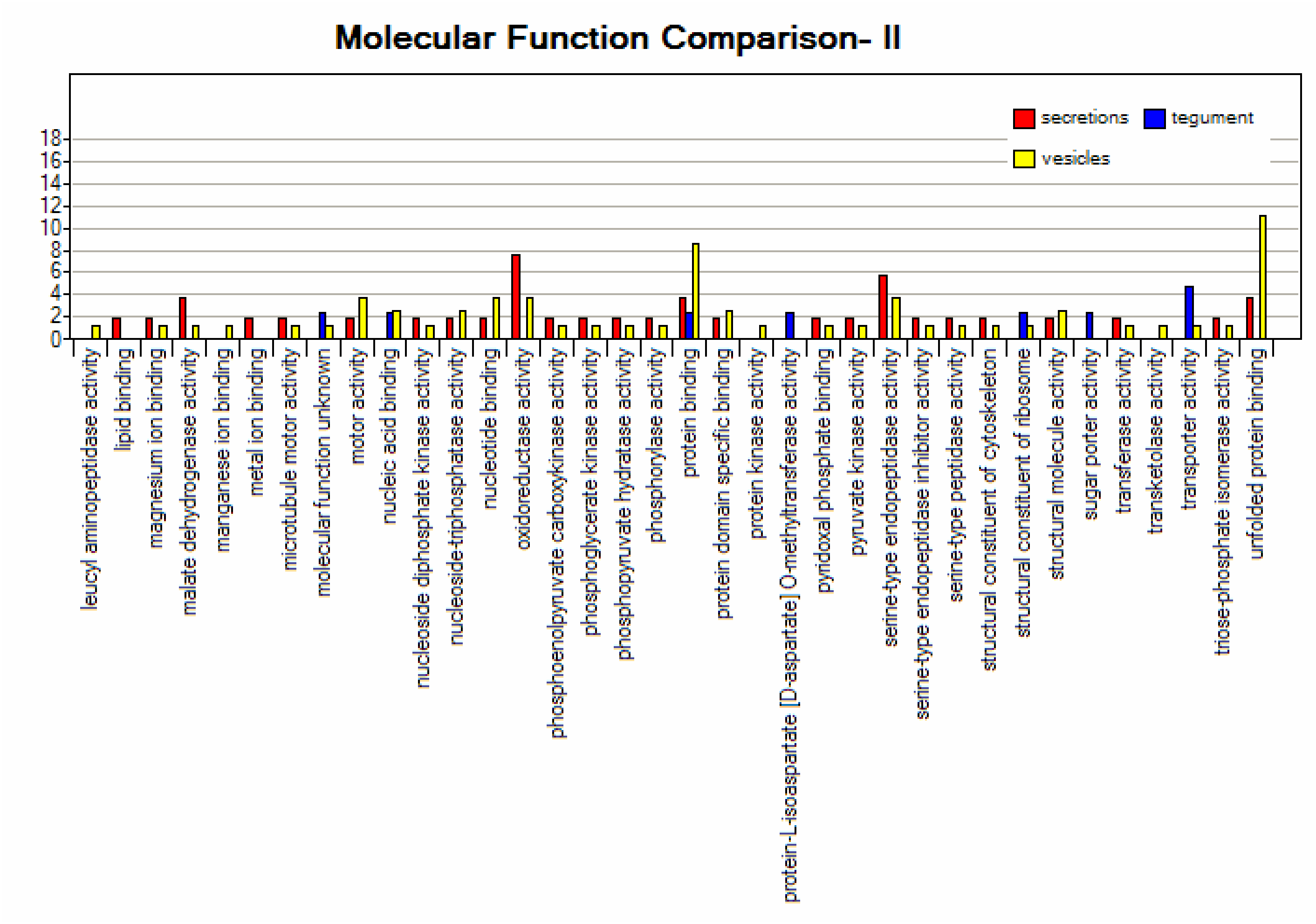


Figure 37. % of sequences corresponding to GO annotations from category molecular function (II)

The control samples have regulation of transcription (7.10%), protein amino acid phosphorylation (5.81%), metabolism (5.16%), cation transport (4.52%), and transport (4.52%) as the most represented terms under the Biological Process category. Under the category Molecular Function, the most represented terms include: ATP binding at 11.61%, DNA binding at 7.74%, catalytic activity and protein kinase activity both at 5.81%, and calcium ion binding at 5.16%. Membrane (11.61%), integral to membrane (10.32%), and nucleus (7.74%) were the most represented terms under the Cellular Component category.

Appendix H shows the entropy calculations that were performed to determine the most significant GO terms. These calculations determine the extent to which each GO term contributes to the predicted class label. The class label in this case is vesicles, secretions, tegument, or control. In comparing vesicles vs. secretions and vesicles vs. tegument ATP binding was found to be the most significant feature/GO term in determining that the protein belonged to the vesicles. Also, between vesicles and tegument, in the Cellular Component category integral to membrane was the most significant feature that determined membership in the class of tegument proteins. Between secretions and tegument, glycolysis was significant in membership in the class of secretion proteins and the cellular component membrane was significant in determining membership in the class of tegument proteins. In comparing the control samples to the vesicles and secretions samples, glycolysis was the most significant term determining membership in either vesicles or secretions, and the cellular component membrane was most significant in determining membership in the class of control proteins. In comparing control proteins to tegument proteins, transporter activity and protein modification were significant in determining membership in the class of tegument proteins and DNA binding and nucleus were significant in determining membership in the class of control proteins.

These results are summarized in Table 9 below. The values in the table indicate GO terms which are significant in determining whether a sample belongs to one sample set versus another. The values ‘V’, ‘S’, and ‘T’ in parentheses indicate which sample class the term favors.

Table 9. GO terms with highest entropy

	V	S	T	C
V		Molecular Function: ATP binding (V)	Molecular Function: ATP binding (V); Cellular Location: integral to membrane (T)	Biological Process: glycolysis (V)
S	Molecular Function: ATP binding (V)		Biological Process: glycolysis (S); Cellular Location: membrane (T)	Biological Process: glycolysis (S); Cellular Location: membrane (T)

T	Molecular Function: ATP binding (V);
	Cellular Location: integral to membrane (T)
		Biological Process: glycolysis (S);
		Cellular Location: membrane (T)
				Molecular Function: transporter activity (T);
				Biological Process: protein modification (T);
				Molecular Function: DNA binding (C);
				Cellular Location: nucleus (C)

C	Biological Process: glycolysis (V)	Biological Process: glycolysis (S);
		Cellular Location: membrane (T)
			Molecular Function: transporter activity (T);
			Biological Process: protein modification (T);
			Molecular Function: DNA binding (C);
			Cellular Location: nucleus (C)	

Protein Domain Analysis

Detailed protein domain annotation data are provided in Appendix G and a visual representation of the data is provided in Figure 38.

In the vesicle samples, three different chaperonin domains are all highly represented (4.94-7.41%) as well as an ATPase domain, two EF-hand domains, and a thioredoxin fold domain, while peptidase domains are represented at about 3.7%. In the secretions, the EF-hand domain is represented in 9.43% of proteins and the thioredoxin fold is represented in 7.55% of proteins, followed by peptidases, dehyrogenase, and histone fold and core all represented in 5.66% of proteins. This corresponds to the GO term percentages that showed that secretions had higher representation of enzymatic and catalytic activity. The

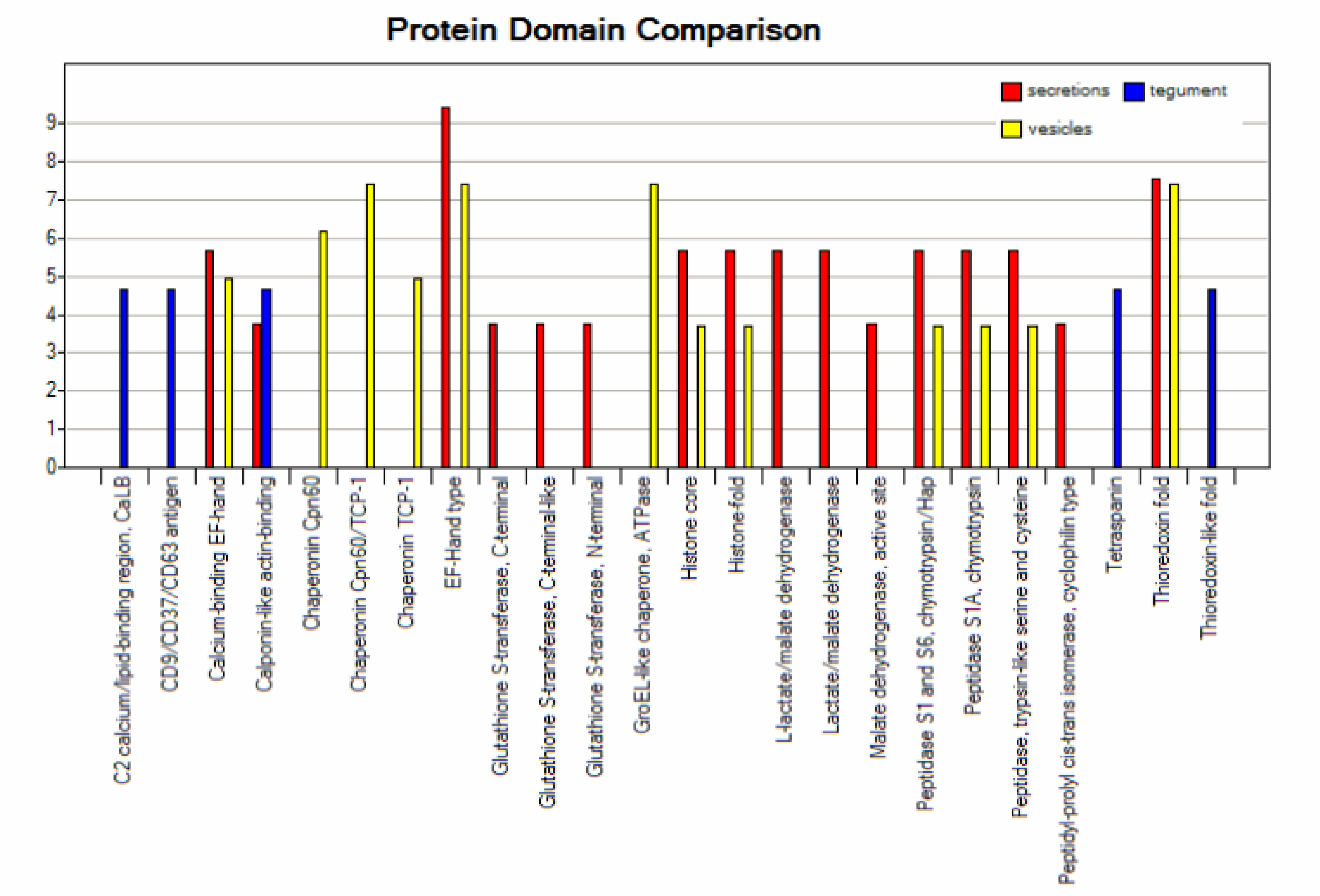


Figure 38. % of sequences corresponding to InterPro protein domains
Only showing those protein domains represented in > 3% of sequences

tegument samples showed calcium/lipid-binding, thioredoxin-like fold, antigen, tetraspanin, and calponin-like actin-binding being equally represented. The control samples indicated protein kinase and protein kinase-like domains being represented very highly (18.60-20.93%), as well as thioredoxin fold, EF-hand type, and hydrolase.

Chaperonin domains are significant in classifying a protein as belonging to vesicles over controls, secretions, and tegument: each of four different chaperonin domains contributes to classifying a protein as being in the vesicles over the controls, chaperonin TCP-1 is also significant in classifying a protein as belonging to vesicles over secretions, and chaperonin Cpn60/TCP-1 in classifying a protein as belonging to vesicles over tegument. The domains histone core, lactate/malate dehydrogenase, and L-lactate/malate dehydrogenase all contribute equally to classifying a protein as belonging to secretions over controls. Calcium/lipid-binding region is somewhat significant in classifying a protein as belonging to the tegument set over the control set. Calcium/lipid-binding region, CD9/CD37/CD63 antigen, and tetraspanin all contribute to classifying a protein as tegument over vesicles. Similarly, calcium/lipid-binding region, CD9/CD37/CD63 antigen, tetraspanin, *and* thioredoxin-like fold contribute to classifying a protein as belonging to tegument over secretions.

These results are summarized in Table 10 below. The values in the table indicate protein domains which are significant in determining whether a sample belongs to one sample set vs. another. The values ‘V’, ‘S’, and ‘T’ in parentheses indicate which class/sample set the domain favors.

Table 10. Protein domains with highest entropy values

	V	**S**	**T**	**C**
V		chaperonin TCP-1 (V)	Calcium/lipid-binding region (T) CD9/CD37/CD63 antigen (T) tetraspanin (T)	Chaperonin Cpn60/TCP-1 (V) GroEL-like chaperone, ATPase (V) Chaperonin Cpn60 (V) Chaperonin TCP-1 (V)
S	chaperonin TCP-1 (V)		calcium/lipid-binding region (T) CD9/CD37/CD63 antigen (T) tetraspanin (T) thioredoxin-like fold (T)	histone core (S) lactate/malate dehydrogenase (S) L-lactate/malate dehydrogenase (S)
T	Calcium/lipid-binding region (T) CD9/CD37/CD63 antigen (T) tetraspanin (T)	calcium/lipid-binding region (T) CD9/CD37/CD63 antigen (T) tetraspanin (T) thioredoxin-like fold (T)		Calcium/lipid-binding region (T)
C	Chaperonin Cpn60/TCP-1 (V) GroEL-like chaperone, ATPase (V) Chaperonin Cpn60 (V) Chaperonin TCP-1 (V)	histone core (S) lactate/malate dehydrogenase (S) L-lactate/malate dehydrogenase (S)	Calcium/lipid-binding region (T)	

D. Summary

A total of 177 proteins found in *S. mansoni* cercarial secretions and adult outer tegumental membrane were analyzed for the presence of conserved motifs using MEME and MAST. A 5-amino acid motif was enriched in the vesicles and found across a total of 11 sample proteins and 2 proteins from an orthologous species, *S. japonicum.* A secretion model was proposed based on functional analysis of the motif and the 13 proteins found to contain it. Sample sets were analyzed for homology using BLAST and CLUSTALW and were annotated with protein domains and GO terms using InterProScan. Methods were developed using Perl, BioPerl, and Matlab for processing InterProScan raw data and determining the most represented, as well as the most statistically significant motifs and GO terms. This data was used to theorize about the effects of the lipid induction versus tail shearing induction experimental methods as well as to analyze the proteins containing the proposed conserved motif.

V. CONCLUSION

A. Overview of Significant Findings

A five amino acid motif, x[K/R]xGE, was found across 12 sample proteins including calreticulin, myosin light chain, calcium binding protein, endoplasmin, protein disulfide isomerase, thioredoxin, thioredoxin peroxidase 1a, thioredoxin 2, and elastases 1a, 1b, and 2a. The mechanically-induced sample was enriched for the motif. It was also found in immunophilin and AUT1 in the orthologous species, *S. japonicum*. These proteins represent three major functions: calcium binding, chaperone (protein folding) activity, and protease activity. The first position of the motif is nonconserved, in the second position is the basic residue lysine or arginine (arginine being specific to the elastases), the third position is nonconserved, in the fourth position is the nonpolar residue glycine, and in the fifth position is the acidic residue glutamic acid. The motif was found at the N-terminal between amino acid 27 and 110, except in the thioredoxin peroxidases, where it was found close to the C-terminal. It was found downstream from a signal peptide in 7 of the 12 sample proteins.

A surprising finding was that the overall percentage of proteins within each sample set that were predicted to contain a signal peptide was significantly lower than the percentage of the control set found to contain a signal peptide. 11% of the vesicles, 9% of the secretions, and 7% of the tegument proteins were predicted to contain a signal peptide in comparison to 23% of control proteins.

The automated workflow developed during this thesis study provides a reliable method for determining which protein domains are most significant. The functional annotation generated in this thesis study correlates well with conclusions reached in the proteomic study of cercarial secretions conducted by Knudsen *et al*., including: 1) proteins related to calcium-binding and regulation were enriched in the mechanically-induced, as well as in the lipid-induced secretion samples, and 2) EF-hand motifs were present in many of the vesicle proteins (Knudsen, Medzihradszky *et al.* 2005). The Molecular Function Comparison I chart in Figure 36 shows that calcium-binding is one of the most represented GO terms in both the vesicles and secretion samples, and the Protein Domain Comparison chart in Figure 38 shows that the EF-hand motif is one of the most represented in the vesicle and secretion samples. This thesis analysis also revealed some unexpected differences between the secretion and vesicle proteins: 1) the biological process protein folding is almost 5% higher in the vesicles than secretions, 2) the molecular function ATP binding is about 12.5% higher in the vesicles than in the secretions, and 3) the biological process glycolysis is almost 10% higher in the secretions than in the vesicles. Entropy analysis revealed that the molecular function ATP binding and several chaperone domains were significant in classifying proteins as belonging to the vesicles over the secretions and the controls. Analysis of the proteins found to be unique to either the vesicle or secretion samples resulted in motor activity and microtubule-based movement being unique to the vesicles and some soluble proteins being unique to the secretions. Possible explanations for this are discussed in the next section.

B. Consideration of Findings in Context of Current Knowledge

The motif was found to not be any of the following: a phosphorylation site, glycosylation site, protease recognition motif, elastase recognition motif, or ER-type signal peptide. It is proposed to be a pro-domain that signals proteins to be shuttled off into special secretory vesicles from the Golgi apparatus. It may be a PC recognition site, as PCs are known to cleave on the C-terminal side of sites composed of single or paired basic amino acids e.g. lysine or arginine (K/R). It bears similarity to the motif found in the malarian parasite, which was experimentally proven to be essential for the export of 320 soluble and insoluble proteins. Similarly to the *Schistosoma* proteins containing the x[K/R]xGE motif, not all of the *Plasmodium* (pfEMP1) proteins found to contain a conserved motif had a predicted signal peptide.

A possible explanation for the low percentage of signal peptides in the adult tegument sample is that the quality of the tegument sequences is lower than the vesicle and secretion samples, resulting in very short sequences in many cases where the signal peptide, if it existed, is missing. Signal peptides may be less frequent in the vesicle and secretion samples because they come from *Schistosoma* cercaria and, as discussed on page 25, cercaria may have evolved sophisticated invasion mechanisms that do not always employ the traditional eukaryotic signal peptides with which SignalP has been trained. The control sample contains proteins from all stages of *S. mansoni* and has a higher percentage of signal peptides. This may be because the life cycle stages other than cercaria use more traditional signal peptides. Also, GO analysis of the control sample shows that 11.61% and 10.32% of proteins belong to the Cellular Location categories membrane and integral to membrane respectively. This is in contrast to the vesicle and secretion samples, neither of which have membrane or integral to membrane as one of the most represented cellular locations; only 3.7% of vesicle proteins have a cellular location of membrane and only 1.89% of secretion proteins had a cellular location of integral to membrane. Perhaps *S. mansoni* employs more traditional signal peptides for membrane proteins and less traditional signal peptides for proteins that are secreted extracellularly. As Bendtsen *et al.*

points out, it is important to note that a negative SignalP prediction does not necessarily mean that the protein is not secreted as some are exported via “non-classical and leaderless pathways” (Bendtsen, Nielsen et al. 2004).

The protein domains and GO terms found to be significant or unique to the vesicle or secretion proteins are likely to be the result of the different experimental sample collection methods. The lipid-induced sample was collected by centrifuging to separate out cercarial bodies and tails from the secretions, while the mechanically-induced sample was collected by scraping vesicles from the surface of the plate. The first method may have caused some of the vesicular membranes, in which some of the chaperone-related proteins may be inserted, to be separated out from the sample along with the cercarial tails and bodies. It may have also resulted in the collection of non-vesicle-specific proteins, such as the cytosolic contents of the acetabular glands. The latter method likely retained most of the vesicular membrane and not as many non-vesicle-specific proteins, resulting in a cleaner sample.

C. Theoretical Implications of the Findings

If the proposed motif can be experimentally validated as necessary for certain proteins to be exported, it could be used to predict secreted proteins. As the motif is present on proteins that are directly exposed to the host and likely to contribute to virulence, it could potentially be used as a target against which a preventative topical drug treatment could be designed. Antigenic variation within proteins makes it difficult to target specific proteins, but if a common motif exists across multiple proteins, a drug could be designed to target/interact with that motif and have activity against multiple virulence factors (the proteins that contain the motif) at once. The drug could be specifically designed to interfere with one or more of the three main functions identified within the proteins found to contain the motif: calcium ion binding, protein folding, or serine protease activity.

The differences between the proteins isolated in the vesicle and secretion samples provide insight into the effects of the different experimental sample preparation methods. The mechanical induction method produced a sample that contained proteins that were more specific to cercarial secretions than the lipid-induced method.

VI. DISCUSSION

A. Limitations of the Study

This study was limited in that it was not complimented by laboratory investigation which could verify that the motif is indeed essential for protein export via secretion vesicles. Also, this study was based on qualitative rather than quantitative proteomic data, which could provide information about which proteins are up and down regulated in secretions. The relatively small sample sizes made it difficult to achieve good MEME e-values. This in turn meant only smaller MAST databases could be searched using the motif revealed by MEME, resulting in larger MAST e-values. A larger sample size would improve e-values of both the MEME and MAST searches. A larger sample size would also allow clustering of the samples based on all sequence features (GO terms and protein domains). Other limitations include the fact that SignalP was not trained on any *S. mansoni* sequences and the fact that GO annotations are not available for every protein domain.

B. Recommendations for Further Research

The proposed conserved motif elucidated in this thesis study needs to be experimentally validated. The ability to transfect *Schistosoma* has been established (Yuan, Shen et al. 2005), as well as the ability to express *Schistosoma* proteins in a heterologous system (Price, Doenhoff et al. 1997). A mutation could be introduced into the proposed motif in the 11 proteins found to contain it and then a reporter gene could be used to track whether the proteins are secreted or not. If the mutation is shown to prevent secretion, then a recombinant expression system could be used to produce large amounts of the proteins as potential antigens for immunological studies.

Additional MEME and MAST iterations should be performed to increase the understanding of the proposed motif or elucidate other conserved motifs. New MAST databases for motif searching could be developed using the larger number of public sequences that are becoming available. All publicly available *Schistosoma* sequences that are annotated with cellular location and life cycle stage could be downloaded, subdivided according to life cycle stage and/or cellular location, and analyzed to search for corresponding conserved motifs and significant domains. The sensitivity of the neural network behind SignalP could be specifically trained on *Schistosoma* proteins to improve its sensitivity and specificity.

The methods presented here for automated analysis of InterProScan raw data could be enhanced in a variety of ways. The Perl and Matlab scripts could be built into an automated workflow and presented as a web application that allows the end-user to submit raw data from InterProScan to retrieve corresponding domain matrices, GO matrices, and entropy calculations. An option could be built into the application that would allow the end-user to automatically extract a smaller matrix from a larger one, based on sequence identifiers, for purposes of being able to focus on specific sets of sequences and their corresponding domains and GO terms within a data set. This was done manually in this thesis study for proteins that were either unique to a sample set and that contained the proposed motif. Analysis of entropy data could be automated such that the most significant domains and GO terms are automatically presented to the user, perhaps in the table format shown in Tables 9 and 10.

APPENDICES

Appendix A: Summary of Sample Data Sets

Following are tables listing all samples used in this study including cercarial vesicles (induction by tail shearing), secretions (induction by skin lipid) and adult tegument. The three sample sets were obtained as detailed in the Methods- Samples and Subjects section. Column heading of sample sets: "Phylo" = phylogenetic group corresponding to phylograms in Results section on pages 50-52. "S?" indicates whether or not a signal peptide was predicted by SignalP. The last column, "KxGEx Motif?", indicates whether or not the sequence contains the motif found in MEME #3.

SAMPLE SET 1: VESICLES					
Description	**Accession number**	**Phylo**	**length (aa)**	**S?**	**xKxGE Motif?**
Pancreatic elastase precursor (elastase 1a)	A28942	2	264	Y	Y
Calcium-binding protein, fluke	A30792	1a	69	N	N
Heat shock protein 86, fluke (fragment)	A45529	1a	442	N	N
Vaccine-dominant antigen Sm21.7	A45630	1b	184	N	N
Tubulin alpha	A48433	1g	451	N	N
Myosin heavy chain	A59287	1a	1940	N	N
Calreticulin	AAA19024	1g	373	Y	Y
Fimbrin	AAA29882	1f	651	N	N
Calcium-binding protein	AAA29921	1a	154	N	Y
glutathione S-transferase, GST [Schistosoma mansoni, Peptide, 218 aa]	AAB21173	1f	218	N	N
Putative cytosol aminopeptidase	AAB41442	4	521	N	N

SAMPLE SET 1: VESICLES					
Description	**Accession number**	**Phylo**	**length (aa)**	**S?**	**xKxGE Motif?**
Calponin homolog	AAB47536	2	361	N	N
Unknown (serpin)	AAB86571	1c	256	N	N
Actin 2	AAC46966	1f	376	N	N
Elastase (elastase 1b)	AAC46967	2	274	Y	Y
Gynecophoral canal protein	AAC47216	1c	688	N	N
Calcium ATPase 2	AAC72756	1b	1011	N	N
Tegumental protein Sm20.8	AAC79131	1b	181	N	N
Thioredoxin peroxidase 2	AAD17299	1a	185	N	Y
Phosphoenolpyruvate carboxykinase	AAD24794	1d	626	N	N
SPO-1 protein (anti-inflammatory protein 6)	AAD26122	1a	117	Y	N
Myosin light chain	AAD41591	1a	160	N	Y
14-3-3 epsilon isoform	AAF21436	3	249	N	N
Endoplasmin	AAF66929	1a	796	Y	Y
SNaK1 (Na+/K+-ATPase alpha)	AAL09322	1b	1007	N	N
Putative histamine-releasing factor	AAL11633	1d	166	N	N
Thioredoxin	AAL79841	1d	106	N	Y
Elastase 2a	AAM43941	2	263	Y	Y
Heat shock protein HSP60	AAM69406	2	549	N	N
Arginase	AAP94031	1f	364	N	N
Actin-binding/filamin-like protein	AAR26703	1b	984	N	N
70000 molecular weight antigen/hsp70 homolog	CAA28976	1g	619	N	N
Elongation factor 1-alpha	CAA69721	4	465	N	N
Protein-disulfide isomerase homolog	CAA80520	1a	482	Y	Y
Similar to myosin regulatory light chain	CD180182	1f	866	N	N
Paramyosin	P06198	1a	211	N	N
Glutathione S-transferase (Sm, 211 aa)	P09792	1f	354	N	N
Major egg antigen P40	P12812	1g	675	N	N
ATP:guanidino kinase SMC74	P16641	2	338	N	N

SAMPLE SET 1: VESICLES					
Description	**Accession number**	**Phylo**	**length (aa)**	**S?**	**xKxGE Motif?**
GAPDH (major larval surface antigen) (P-37)	P20287	4	416	N	N
Phosphoglycerate kinase	P41759	1f	284	N	N
Tropomyosin 1 (TMI) (polypeptide 49)	P42637	1a	284	N	N
Tropomyosin 2 (TMII)	P42638	1a	253	N	N
Triose-phosphate isomerase (TPI)	P48501	1e	363	N	N
Fructose-bisphosphate aldolase	P53442	2	252	N	N
14-3-3 protein homolog 1	Q26540	3	161	N	N
Peptidyl-prolyl cis-trans isomerase (PPIase)	Q26565	1g	434	N	N
Enolase (2-phosphoglycerate dehydratase)	Q27877	1g	89	N	N
Probable dynein light chain (SM10)	Q94748	1f	545	N	N
T-complex protein-1, alpha subunit	Q94757	2	1679	N	N
Surface protein, fluke	T30271	1d	100	Y	N
Weakly similar to heterogeneous nuclear ribonucleoprotein A2 homolog 1	TC10489	1g	314	N	N
Similar to ATP synthase alpha-chain mito	TC10585	1d	547	N	N
Similar to transitional ER ATPase	TC10596	1g	803	N	N
Weakly similar to phosphoglucomutase	TC10613	2	601	N	N
Similar to citrate synthase	TC10655	1g	433	N	N
Similar to B-cell receptor-associated protein 32	TC10689	1g	274	Y	N
Similar to ribosomal protein L12	TC10765	2	165	N	N
Thioredoxin peroxidase 3	TC10839	1a	219	N	N
Similar to nucleoside-diphosphate kinase	TC11413	1g	149	N	N
Similar to glycogen phosphorylase, muscle	TC13591	1e	847	N	N
Similar to ATP synthase [1]-chain mito	TC13604	1g	517	N	N
Homolog to H2B histone	TC13606	1e	122	N	N
Similar to chaperonin containing T-complex protein-1, beta subunit	TC13620	2	530	N	N
Similar to histone H3	TC13658	4	136	N	N
Similar to chaperonin containing T-complex protein-1, gamma subunit	TC13671	2	466	N	N
Similar to succinate dehydrogenase Fp subunit	TC13974	3	258	N	N

SAMPLE SET 1: VESICLES					
Description	**Accession number**	**Phylo**	**length (aa)**	**S?**	**xKxGE Motif?**
Similar to histone H4	TC14578	2	103	N	N
Similar to transketolase	TC16539	2	631	N	N
Similar to lactate dehydrogenase	TC16735	3	332	N	N
Homolog to calmodulin	TC16812	1a	149	N	N
Similar to malate dehydrogenase precursor	TC16844	3	341	N	N
Similar to ADP/ATP translocase	TC16858	1e	221	N	N
Similar to chaperonin containing T-complex protein-1, zeta subunit	TC16896	2	547	N	N
Similar to hypothetical Schistosoma japonicum protein	TC17017	1g	119	N	N
Similar to T-complex protein-1, epsilon subunit	TC6878	2	545	N	N
Homolog to tubulin beta-2 chain	TC7336	1g	443	N	N
Weakly similar to troponin T	TC7449	1a	325	N	N
Similar to pyruvate kinase	TC7454	1f	572	N	N
Similar to HEL protein	TC7459	2	440	N	N
Phosphoglycerate mutase	TC7546	1d	250	N	N

SAMPLE SET 2: SECRETIONS					
Description	**Accession number**	**Phylo**	**length (aa)**	**S?**	**xKxGE Motif?**
Pancreatic elastase precursor (elastase 1a) (Sm)	A28942	1a	264	Y	Y
Calpain (EC 3.4.22.17) large chain (Sm)	A39343	1a	758	N	N
Heat shock protein 86 (Sm)	A45529	1b	442	N	N
Vaccine-dominant antigen Sm21.7 (Sm)	A45630	2	184	N	N
Fimbrin (Sm)	AAA2988	2	651	N	N
Calcium-binding protein (Sm)	AAA29921	1a	154	N	Y
GST (Sm, 218 aa)	AAB21173	3	218	N	N
Putative cytosol aminopeptidase (Sm)	AAB41442	1a	520	N	N
Calponin homolog (Sm)	AAB47536	1c	361	N	N

SAMPLE SET 2: SECRETIONS					
Description	**Accession number**	**Phylo**	**length (aa)**	**S?**	**xKxGE Motif?**
Unknown (serpin) (Sm)	AAB86571	3	256	N	N
Cu,Zn-superoxide dismutase (Sm)	AAC14467	1b	153	N	N
Similar to carbonyl reductase (Sm)	AAC46898	1c	276	N	N
Actin 2 (Sm)	AAC46966	3	376	N	N
Elastase (elastase 1b) (Sm)	AAC46967	1a	274	Y	Y
Tegumental protein Sm20.8 (Sm)	AAC79131	2	181	N	N
Thioredoxin peroxidase 2 (Sm)	AAD17299	1c	185	N	Y
Phosphoenolpyruvate carboxykinase (Sm)	AAD24794	2	626	N	N
SPO-1 protein (anti-inflammatory protein 6) (Sm)	AAD26122	3	117	Y	N
Fatty acid-binding protein Sm14 (Sm)	AAL15461	1a	133	N	N
Thioredoxin (Sm)	AAL79841	3	106	N	Y
Elastase 2a (Sm)	AAM43941	1a	263	Y	Y
Heat shock protein HSP60 (Sm)	AAM69406	1b	549	N	N
ATP-diphosphohydrolase 1 (Sm)	AAP94734	1a	544	N	N
Cysteine protease inhibitor (Sm)	AAQ16180	1c	101	N	N
Actin-binding/filamin-like protein (Sm)	AAR26703	3	984	N	N
70,000 molecular weight antigen/hsp70 homolog (Sm)	CAA28976	1a	619	N	N
Elongation factor 1 (Sm)	CAA69721	1a	465	N	N
Glutathione *S*-transferase, 28 kDa (GST 28) (Sm)	P09792	3	211	N	N
ATP:guanidino kinase SMC74 (Sm)	P16641	1c	675	N	N
GAPDH (major larval surface antigen) (P-37) (Sm)	P20287	1b	338	N	N
Ferritin-2 heavy chain (Sm)	P25319	3	173	N	N
Phosphoglycerate kinase (Sm)	P41759	1b	416	N	N
Triose-phosphate isomerase (TIM) (Sm)	P48501	2	253	N	N
Fructose-bisphosphate aldolase (Sm)	P53442	2	363	N	N
14-3-3 protein homolog 1 (Sm)	Q26540	2	252	N	N
Peptidyl-prolyl cis-trans isomerase B precursor (Sm)	Q26551	1b	213	Y	N
Peptidyl-prolyl cis-trans isomerase (PPIase) (Sm) 1	Q26565	1b	161	N	N

SAMPLE SET 2: SECRETIONS					
Description	**Accession number**	**Phylo**	**length (aa)**	**S?**	**xKxGE Motif?**
6-Phosphofructokinase (Sm)	Q27778	2	781	N	N
Enolase (2-phosphoglycerate dehydratase) (Sm)	Q27877	1b	434	N	N
Probable dynein light chain (SM10) (T-cell-stimulating antigen SM10)	Q94748	3	89	N	N
Similar to nucleoside-diphosphate kinase (Sm)	TC11413	1c	149	N	N
Similar to muscle glycogen phosphorylase (Sm)	TC13591	1a	847	N	N
Similar to ATP synthase _-chain mito (Sm)	TC13604	3	517	N	N
Homolog to H2B histone (Sm)	TC13606	1b	122	N	N
Similar to histone H3 (Sm)	TC13658	1a	136	N	N
Similar to histone H4 (Sm)	TC14578	1a	103	N	N
Weakly similar to lactate dehydrogenase (Sm)	TC16735	1c	332	N	N
Homolog to calmodulin (Sm)	TC16812	1a	149	N	N
Similar to malate dehydrogenase, mito (Sm)	TC16844	2	341	N	N
Similar to malate dehydrogenase, cytosolic (Sm)	TC17066	2	329	N	N
Homolog to tubulin _-2 chain (Sm)	TC7336	1a	443	N	N
Similar to pyruvate kinase (Sm)	TC7454	1a	572	N	N
Homolog to phosphoglycerate mutase (Sm)	TC7546	2	250	N	N

SAMPLE SET 3: TEGUMENT					
Description	**Accession number**	**Phylo**	**length (aa)**	**S?**	**xKxGE Motif?**
no significant homology found	AI882660	3	99	N	N
no significant homology found	BF936329	1b	31	N	N
annexin 11a isoform 2	CD098410	1b	90	N	N
Amidase	CD156396	1c	59	N	N
Dysferlin	CD157335	1b	71	N	N
LIM and SH3 protein 1b	CD182193	1a	75	N	N
no significant homology found	CD192195	1a	79	N	N

SAMPLE SET 3: TEGUMENT					
Description	**Accession number**	**Phylo**	**length (aa)**	**S?**	**xKxGE Motif?**
DEAD (Asp-Glu-Ala-Asp) box polypeptide 1	CD194580	2	47	N	N
SCP-like extracellular proteinb	TC10634	1a	189	N	N
SCP-like extracellular proteinb	TC10635	1a	303	N	N
aquaporin-3	TC10637	1a	304	N	N
22K surface membrane antigen	TC10917	1b	182	N	N
similar to Pcmt1-prov protein	TC11206	1b	237	N	N
Exocyst complex component Sec10 (hSec10)	TC11347	3	267	N	N
ATP-diphosphohydrolase 1a	TC11432	2	545	N	N
putative related to F3G5b	TC11571	1b	203	N	N
ubiquitin/ribosomal fusion proteinb	TC11590	1b	128	N	N
phosphatidylinositol transfer protein	TC12230	2	119	N	N
no significant homology found	TC13203	1a	91	N	N
Nebulette	TC13379	3	47	Y	N
breast adenocarcinoma marker like (26.0 kD) (2N58)	TC13662	3	237	N	N
CD63-like protein Sm-TSP-2a	TC13732	1c	219	Y	N
Myosin D	TC13851	3	274	N	N
Alpha-actinin, sarcomeric (F-actin cross linking protein)	TC14047	3	288	N	N
thioredoxin peroxidase 1a	TC14049	1c	185	N	Y
no significant homology found	TC14238	1a	277	N	N
ectonucleotide pyrophosphatase/phosphodiesterase 5	TC14339	3	271	N	N
syntenin	TC14697	1a	225	N	N
filamin isoform A	TC17006	3	472	N	N
SGTP4	TC17442	2	458	N	N
fatty acid coenzyme A ligase 5	TC17495	1c	262	N	N
no significant homology found	TC18339	1c	236	N	N
no significant homology found	TC19226	2	165	N	N
HLA-B associated transcript 1b	TC7459	1c	440	N	N
major egg antigen	TC7485	1a	366	N	N

SAMPLE SET 3: TEGUMENT					
Description	**Accession number**	**Phylo**	**length (aa)**	**S?**	**xKxGE Motif?**
fimbrin	TC7585	1a	651	N	N
rat coatamer beta subunit	TC7811	1b	460	N	N
hypothetical proteinb	TC7948	1a	394	N	N
actin-binding and severin family group-protein	TC8208	1a	68	N	N
adenylyl cyclase-associated proteinb	TC8265	2	234	N	N
no significant homology found	TC8556	1a	203	Y	N
no significant homology found	TC9174	1c	55	N	N
no significant homology found	TC9780	3	37	N	N

Appendix B: Control Data Set

This data set was obtained as described in Methods- Samples and Subjects.

155 Background/Control Sequences (source: NCBI's nonredundant protein database)

accession number	description	Length
>gi_1002620_gb_AAC46888.1_	unknown	98
>gi_1002624_gb_AAC46890.1_	similar to synaptobrevin	102
>gi_1002666_gb_AAC46893.1_	unknown	78
>gi_1002668_gb_AAC46894.1_	unknown	239
>gi_1002670_gb_AAC46895.1_	unknown	236
>gi_1002674_gb_AAC46897.1_	similar to E. coli phosphoserine phosphohydrolase	223
>gi_1002676_gb_AAC46898.1_	similar to human carbonyl reductase (NADPH)	276
>gi_1002682_gb_AAC46900.1_	similar to mitochondrial ATPase inhibitor	63
>gi_1002684_gb_AAC46901.1_	similar to synaptobrevin	54
>gi_1016750_gb_AAA79138.1_	rab-related GTP-binding protein	205
>gi_10281265_gb_AAG15509.1_AF301004_1	thioredoxin peroxidase 3 [Sm]	219
>gi_10442652_gb_AAG17406.1_AF283511_1	receptor kinase I-interacting protein SIP [Sm]	455
>gi_1090775_prf__2019440B	Sm65 antigen	369
>gi_10953801_gb_AAG25600.1_AF297468_1	chromatin assembly factor 1 small subunit-like protein [Sm]	308
>gi_11464653_gb_AAG35265.1_AF215933_1	Smad1 [Sm]	455
>gi_115391_sp_P13566_CABP_SCHMA	Calcium-binding protein (CaBP)	69
>gi_11596371_gb_AAG38588.1_AF316828_1	zinc finger protein SmZF1 [Sm]	164
>gi_119756_sp_P16463_F801_SCHMA	Female specific 800 protein (FS800)	238
>gi_12248339_gb_AAG13165.2_	NADH dehydrogenase subunit 4L [Sm]	86
>gi_12249163_ref_NP_066213.2_	NADH dehydrogenase subunit 4L [Sm]	86
>gi_12958632_gb_AAK09382.1_AF321922_1	purine-nucleoside phosphorylase [Sm]	287
>gi_1354127_gb_AAC47216.1_	gynecophoral canal protein	688
>gi_13958030_gb_AAK50768.1_AF361357_1	Ca-ATPase-like protein SMA3 [Sm]	1035
>gi_14192680_gb_AAC16404.3_	receptor kinase-1 precursor [Sm]	780

155 Background/Control Sequences (source: NCBI's nonredundant protein database)

accession number	description	Length
>gi_1449406_gb_AAC37263.1_	NF-YA subunit	268
>gi_15127836_gb_AAK84311.1_AF361883_1	high voltage-activated calcium channel Cav2A [Sm]	2203
>gi_15127838_gb_AAK84312.1_AF361884_1	high voltage-activated calcium channel Cav1 [Sm]	1776
>gi_15149312_gb_AAK85233.1_AF395822_1	thioredoxin glutathione reductase [Sm]	598
>gi_15420528_gb_AAK97376.1_AF358445_1	glutaminyl-tRNA synthetase [Sm]	531
>gi_1575028_gb_AAB09439.1_	Psmras1 [Sm]	184
>gi_1580810_emb_CAA67208.1_	T-cell-stimulating antigen [Sm]	89
>gi_15808978_gb_AAL08579.1_AF418550_1	albumin precursor [Sm]	608
>gi_15824396_gb_AAL09322.1_AF303222_1	SNaK1 [Sm]	1007
>gi_15986653_gb_AAL11699.1_AF375996_1	14-3-3 epsilon 2 [Sm]	249
>gi_15986655_gb_AAL11700.1_AF376135_1	eukaryotic translation initiation factor 2 alpha subunit [Sm]	328
>gi_160941_gb_AAA29861.1_	calcium binding protein	69
>gi_161050_gb_AAA29908.1_	p48 eggshell protein	414
>gi_161052_gb_AAA29910.1_	ORF 2	382
>gi_1619614_emb_CAA69721.1_	elongation factor 1-alpha [Sm]	465
>gi_1620592_gb_AAC47307.1_	dynein light chain	89
>gi_16876479_gb_AAF21638.2_AF031196_1	histamine-responsive G-protein coupled receptor [Sm]	560
>gi_17907114_emb_CAD13249.1_	tyrosine kinase [Sm]	1264
>gi_18181863_emb_CAC85211.2_	cathepsin B endopeptidase [Sm]	347
>gi_18369833_gb_AAL67949.1_	receptor tyrosine kinase [Sm]	1559
>gi_1841843_gb_AAB47536.1_	calponin homolog [Sm]	361
>gi_18874552_gb_AAL79841.1_AF473536_1	thioredoxin [Sm]	106
>gi_19071249_gb_AAL84173.1_AF422164_1	receptor for activated PKC [Sm]	315
>gi_20270936_gb_AAM18481.1_AF492390_1	Sm14 fatty acid-binding protein delta E3 variant [Sm]	98
>gi_20384923_gb_AAM09083.1_	Na+/Cl- dependent neurotransmitter transporter-like protein [Sm]	761
>gi_21217531_gb_AAM43941.1_AF510339_1	elastase 2a [Sm]	263
>gi_21217533_gb_AAM43942.1_AF510340_1	elastase 2b [Sm]	263

155 Background/Control Sequences (source: NCBI's nonredundant protein database)

accession number	description	Length
>gi_2131129_emb_CAA73329.1_	amidase [Sm]	691
>gi_21436485_gb_AAM51567.1_	immunophilin FK506 binding protein FKBP12 [Sm]	108
>gi_22074178_gb_AAK98796.1_	ferredoxin NADP+ reductase [Sm]	522
>gi_2246652_gb_AAC62254.1_	lysophospholipase homolog [Sm]	239
>gi_22531389_emb_CAD44625.1_	cathepsin B1 isotype 2 [Sm]	340
>gi_23094378_emb_CAB93676.2_	calcineurin A [Sm]	644
>gi_23305770_gb_AAN17275.1_	unknown [Sm]	114
>gi_23305772_gb_AAN17276.1_	CD63-like protein Sm-TSP-2 [Sm]	239
>gi_23305778_gb_AAN17279.1_	unknown [Sm]	156
>gi_2345100_gb_AAC02298.1_	Pad1 homolog [Sm]	313
>gi_2345102_gb_AAC02299.1_	trans-spliced variant protein [Sm]	167
>gi_23663956_gb_AAN39120.1_AF314754_1	insulin receptor protein kinase RTK-2 [Sm]	1499
>gi_24415108_gb_AAN59790.1_	trimeric G-protein alpha o subunit [Sm]	328
>gi_24415111_gb_AAN59791.1_	trimeric G-protein gamma subunit [Sm]	66
>gi_2494322_sp_Q26571_IF5A_SCHMA	Eukaryotic translation initiation factor 5A-2 (eIF-5A)	52
>gi_2623840_gb_AAB86568.1_	unknown [Sm]	406
>gi_26245438_gb_AAN77581.1_	Rho3 GTPase [Sm]	242
>gi_2665824_gb_AAB88508.1_	ribosomal protein L37 [Sm]	88
>gi_27657926_gb_AAO18222.1_	peptidylglycine alpha hydroxylating mono-oxygenase [Sm]	350
>gi_27699497_gb_AAN17278.2_	CD9-like protein Sm-TSP-1 [Sm]	247
>gi_28192573_gb_AAO21365.1_	leucine-rich protein [Sm]	296
>gi_28261409_tpg_DAA00890.1_	TPA: reverse transcriptase [Sm]	789
>gi_2842736_sp_Q94748_DYL2_SCHMA	Probable dynein light chain (T-cell-stimulating antigen SM10)	89
>gi_2842737_sp_Q94758_DYL1_SCHMA	Dynein light chain	89
>gi_28628851_gb_AAO49385.1_AF484940_1	glutathione S-transferase omega [Sm]	241
>gi_30142122_gb_AAP13803.1_	developmentally regulated antigen 10.3 precursor [Sm]	200
>gi_31746497_gb_AAP68901.1_AF515706_1	Fes-like tyrosine kinase protein [Sm]	1259
>gi_3282676_gb_AAC28780.1_	nuclear factor Y transcription factor subunit B homolog [Sm]	242

155 Background/Control Sequences (source: NCBI's nonredundant protein database)

accession number	description	Length
>gi_3283986_gb_AAC25419.1_	13 kDa tegumental antigen Sm13 [Sm]	104
>gi_33089916_gb_AAP93838.1_	tyrosinase 1 precursor [Sm]	481
>gi_33114187_gb_AAP94734.1_	ATP-diphosphohydrolase 1 [Sm]	544
>gi_33242492_gb_AAQ00945.1_	general control nonrepressed 5 [Sm]	899
>gi_33339659_gb_AAQ14321.1_	acetylcholinesterase [Sm]	687
>gi_33355623_gb_AAQ16180.1_	cysteine protease inhibitor [Sm]	101
>gi_3337405_gb_AAC27440.1_	alpha-(1 3)-fucosyltransferase VII; SmFuct [Sm]	351
>gi_34099845_gb_AAQ57211.1_	putative neuropeptide receptor [Sm]	492
>gi_35187018_gb_AAQ84177.1_	Smad4 [Sm]	738
>gi_37029992_gb_AAQ88098.1_	hox protein Dfd [Sm]	543
>gi_37702159_gb_AAR00731.1_	protein kinase C type beta [Sm]	618
>gi_37722427_gb_AAN72832.1_	LOK-like protein kinase [Sm]	1056
>gi_37776869_emb_CAE51198.1_	src tyrosine kinase [Sm]	647
>gi_37982950_gb_AAR06260.1_	heat shock transcription factor [Sm]	154
>gi_38004412_gb_AAR07505.1_	Abd-A-like protein [Sm]	718
>gi_3859490_gb_AAC72756.1_	calcium ATPase 2 [Sm]	1011
>gi_38683290_gb_AAR26703.1_	actin-binding/filamin-like protein [Sm]	984
>gi_3907627_gb_AAC78683.1_	G-box binding factor homolog [Sm]	214
>gi_40317595_gb_AAR84361.1_	nicotinic acetylcholine receptor alpha subunit precursor [Sm]	686
>gi_40317597_gb_AAR84362.1_	nicotinic acetylcholine receptor non-alpha subunit precursor [Sm]	731
>gi_40365359_gb_AAR85353.1_	high mobility group B1 protein [Sm]	176
>gi_40365361_gb_AAR85354.1_	ubiquitin [Sm]	103
>gi_407047_gb_AAB39265.1_	amino acid permease	503
>gi_40743702_gb_AAR89512.1_	ankyrin-like protein [Sm]	181
>gi_4090941_gb_AAC98911.1_	Sm29 [Sm]	191
>gi_4104017_gb_AAD01923.1_	tryptophan hydroxylase; SmTPH [Sm]	497
>gi_41393747_gb_AAP94031.1_	arginase [Sm]	364

155 Background/Control Sequences (source: NCBI's nonredundant protein database)

accession number	description	Length
>gi_42405318_gb_AAS13487.1_	Ftz-F1 interacting protein [Sm]	788
>gi_425474_gb_AAA66476.1_	SMDR1	691
>gi_4322668_gb_AAD16119.1_	retinoic acid receptor RXR [Sm]	743
>gi_44829165_tpg_DAA04496.1_	TPA: gag protein [Sm]	277
>gi_44829167_tpg_DAA04497.1_	TPA: pol polyprotein [Sm]	1227
>gi_44829169_tpg_DAA04498.1_	TPA: pol polyprotein [Sm]	1680
>gi_44829171_tpg_DAA04499.1_	TPA: pol polyprotein [Sm]	1382
>gi_44829173_tpg_DAA04500.1_	TPA: pol polyprotein [Sm]	1166
>gi_44829174_tpg_DAA04501.1_	TPA: gag protein [Sm]	288
>gi_44829175_tpg_DAA04502.1_	TPA: ORF3 [Sm]	232
>gi_454258_emb_CAA82848.1_	unnamed protein product [Sm]	98
>gi_4581919_gb_AAD24794.1_AF120929_1	phosphoenolpyruvate carboxykinase [Sm]	626
>gi_4753140_gb_AAC79802.3_	annexin [Sm]	365
>gi_477295_pir__A48570	cystatin homolog - fluke (Sm)	89
>gi_48475044_gb_AAR84066.2_	putative seven transmembrane receptor [Sm]	366
>gi_49473446_gb_AAT66412.1_	polyprotein [Sm]	201
>gi_495668_gb_AAA29882.1_	fimbrin	651
>gi_499349_gb_AAA19024.1_	calreticulin	373
>gi_50402589_gb_AAT76629.1_	thioredoxin 2 [Sm]	104
>gi_50429226_gb_AAT77204.1_	neuropeptide F precursor [Sm]	147
>gi_50442714_gb_AAT77263.1_	methionine sulfoxide reductase B2a [Sm]	175
>gi_50442729_gb_AAT77264.1_	methionine sulfoxide reductase B2b [Sm]	137
>gi_510098_gb_AAC37227.1_	potassium channel protein	512
>gi_514912_gb_AAB81008.1_	LGG	196
>gi_51988420_emb_CAH04147.1_	P2X ATP gated ion channel [Sm]	437
>gi_52222500_gb_AAU34080.1_	glutathione peroxidase-2 [Sm]	179
>gi_5305329_gb_AAD41591.1_	myosin light chain [Sm]	160
>gi_542452_pir__S42030	hypothetical protein - fluke (Sm)	98

155 Background/Control Sequences (source: NCBI's nonredundant protein database)

accession number	description	Length
>gi_55139749_gb_AAV41489.1_	Sm50 protein [Sm]	466
>gi_552247_gb_AAA29913.1_	phosphagen kinase [Sm]	52
>gi_55274739_gb_AAV49163.1_	polo-like kinase [Sm]	618
>gi_5566124_gb_AAD45325.1_AF158102_1	retinoid-x-receptor [Sm]	784
>gi_55793504_gb_AAV65746.1_	glycerol 3-phosphate dehydrogenase [Sm]	350
>gi_60172784_gb_AAX14497.1_	hox protein Smox1 [Sm]	745
>gi_61660998_gb_AAX51223.1_	mitochondrial thioredoxin precursor [Sm]	147
>gi_62114973_gb_AAX63737.1_	carbohydrate-binding calcium-dependent lectin precursor [Sm]	409
>gi_6649234_gb_AAF21436.1_AF195529_1	14-3-3 epsilon [Sm]	249
>gi_6650016_gb_AAF21676.1_AF051138_1	trispanning orphan receptor; TORE [Sm]	281
>gi_6841028_gb_AAF28867.1_	cyclophilin [Sm]	181
>gi_732959_emb_CAA88616.1_	eukaryotic translation initiation factor 5A [Sm]	52
>gi_7595982_gb_AAF64527.1_AF254148_1	PUR-alpha-like protein [Sm]	267
>gi_7673568_gb_AAF66929.1_AF217404_1	endoplasmin [Sm]	796
>gi_7960045_gb_AAF71198.1_AF183577_1	alpha 38720 fucosyltransferase [Sm]	426
>gi_808821_gb_AAA96714.1_	ATPase	1022
>gi_8132429_gb_AAF73286.1_AF155134_1	RHO G-protein coupled receptor [Sm]	381
>gi_8250653_emb_CAB93677.1_	calcineurin B [Sm]	169
>gi_9081807_gb_AAF82607.1_	AUT1 [Sm]	349
>gi_951431_gb_AAA74696.1_	ORF-RF2; putative	186
>gi_9828608_gb_AAG00234.1_AF283817_1	dolichol phosphate mannose synthase [Sm]	237

Appendix C: Sequence Homology Within Sample Sets

Vesicle proteins with homology to other vesicle proteins

(More than 50% homology over at least 100 residues with an expectation value of 1x10-5 or better.)

accession	description	length	num hits	hit accession	hit description	hit significance	hsp length	% identity
A28942	pancreatic elastase (EC 3.4.21.36) precursor - fluke (Sm)	264	8	AAC46967	elastase	1.00E-144	275	92.36
AAC46967	elastase	274	4	A28942	pancreatic elastase (EC 3.4.21.36) precursor - fluke (Sm)	1.00E-144	275	92.36
A28942	pancreatic elastase (EC 3.4.21.36) precursor - fluke (Sm)	264	8	AAM43941	elastase 2a [Sm]	1.00E-124	259	82.24
AAM43941	elastase 2a [Sm]	263	6	A28942	pancreatic elastase (EC 3.4.21.36) precursor - fluke (Sm)	1.00E-124	259	82.24
AAC46967	elastase	274	4	AAM43941	elastase 2a [Sm]	1.00E-124	274	79.93
AAM43941	elastase 2a [Sm]	263	6	AAC46967	elastase	1.00E-124	274	79.93
P42637	Tropomyosin 1 (TMI) (Polypeptide 49)	284	17	P42638	Tropomyosin 2 (TMII)	1.00E-101	283	66.08
P42638	Tropomyosin 2 (TMII)	284	13	P42637	Tropomyosin 1 (TMI) (Polypeptide 49)	1.00E-101	283	66.08
AAD17299	thioredoxin peroxidase [Sm]	185	4	TC10839	ORF 71..729 frame +2	7.00E-63	179	59.78
TC10839	ORF 71..729 frame +2	219	2	AAD17299	thioredoxin peroxidase [Sm]	8.00E-63	179	59.78
Q26540	14-3-3 protein homolog 1	252	7	AAF21436	14-3-3 epsilon [Sm]	1.00E-69	239	54.81
AAF21436	14-3-3 epsilon [Sm]	249	10	Q26540	14-3-3 protein homolog 1	1.00E-69	239	54.81

Secretion proteins with homology to other secretion proteins

(More than 50% homology over at least 100 residues with an expectation value of 1x10-5 or better)

accession	description	length	num hits	hit accession	hit description	hit significance	hsp length	% identity
A28942	pancreatic elastase (EC 3.4.21.36) precursor - fluke (Sm)	264	6	AAC46967	elastase	1.00E-144	275	92.36
AAC46967	elastase	274	5	A28942	pancreatic elastase (EC 3.4.21.36) precursor - fluke (Sm)	1.00E-144	275	92.36
A28942	pancreatic elastase (EC 3.4.21.36) precursor - fluke (Sm)	264	6	AAM43941	elastase 2a [Sm]	1.00E-125	259	82.24
AAM43941	elastase 2a [Sm]	263	6	A28942	pancreatic elastase (EC 3.4.21.36) precursor - fluke (Sm)	1.00E-125	259	82.24
AAC46967	elastase	274	5	AAM43941	elastase 2a [Sm]	1.00E-124	274	79.93
AAM43941	elastase 2a [Sm]	263	6	AAC46967	elastase	1.00E-124	274	79.93
Q26551	Peptidyl-prolyl cis-trans isomerase B precursor (PPIase) (Rotamase) (Cyclophilin B) (S-cyclophilin)	213	5	Q26565	Peptidyl-prolyl cis-trans isomerase (PPIase) (Rotamase) (Cyclophilin) (Cyclosporin A-binding protein) (p17.7) (Smp17.7)	1.00E-45	164	56.71
Q26565	Peptidyl-prolyl cis-trans isomerase (PPIase) (Rotamase) (Cyclophilin) (Cyclosporin A-binding protein)	161	6	Q26551	Peptidyl-prolyl cis-trans isomerase B precursor (PPIase) (Rotamase) (Cyclophilin B) (S-cyclophilin)	1.00E-45	164	56.71

Tegument proteins with homology to other tegument proteins

(More than 50% homology over at least 100 residues with an expectation value of 1x10-5 or better)

accession	Description	length	num hits	hit accession	hit description	hit significance	hsp length	% identity
TC10635	TC2046 TC2168 TC558 TC722 TC3699 TC4460 TC6565 frame +1	303	6	TC10634	frame +1	9.00E-42	109	74.31
TC10634	frame +1	189	9	TC10635	TC2046 TC2168 TC558 TC722 TC3699 TC4460 TC6565 frame +1	5.00E-42	109	74.31

Control proteins with homology to other control proteins

(More than 50% homology over at least 100 residues with an expectation value of 1x10-5 or better)

accession	description	length	num hits	hit accession	hit description	hit sig	hsp len	% identity
gi_2345102_gb_AAC02299.1_	trans-spliced variant protein [Sm]	167	4	gi_2345100_gb_AAC02298.1_	Pad1 homolog [Sm]	7.00E-62	121	92.56
gi_2345100_gb_AAC02298.1_	Pad1 homolog [Sm]	313	7	gi_2345102_gb_AAC02299.1_	trans-spliced variant protein [Sm]	2.00E-61	121	92.56
gi_3859490_gb_AAC72756.1_	calcium ATPase 2 [Sm]	1011	8	gi_808821_gb_AAA96714.1_	ATPase	0	1002	69.66
gi_808821_gb_AAA96714.1_	ATPase	1022	7	gi_3859490_gb_AAC72756.1_	calcium ATPase 2 [Sm]	0	1002	69.66
gi_21217533_gb_AAM43942.1_AF510340_1	elastase 2b [Sm]	263	4	gi_21217531_gb_AAM43941.1_AF510339_1	elastase 2a [Sm]	1.00E-88	263	58.56
gi_21217531_gb_AAM43941.1_AF510339_1	elastase 2a [Sm]	263	2	gi_21217533_gb_AAM43942.1_AF510340_1	elastase 2b [Sm]	1.00E-88	263	58.56

accession	description	length	num hits	hit accession	hit description	hit sig	hsp len	% identity
gi_15986653_gb_AAL11699.1_AF375996_1	14-3-3 epsilon 2 [Sm]	249	5	gi_6649234_gb_AAF21436.1_AF195529_1	14-3-3 epsilon [Sm]	8.00E-72	246	57.32
gi_6649234_gb_AAF21436.1_AF195529_1	14-3-3 epsilon [Sm]	249	7	gi_15986653_gb_AAL11699.1_AF375996_1	14-3-3 epsilon 2 [Sm]	8.00E-72	246	57.32
gi_11464653_gb_AAG35265.1_AF215933_1	Smad1 [Sm]	455	11	gi_35187018_gb_AAQ84177.1_	Smad4 [Sm]	9.00E-38	122	52.46
gi_35187018_gb_AAQ84177.1_	Smad4 [Sm]	738	12	gi_11464653_gb_AAG35265.1_AF215933_1	Smad1 [Sm]	1.00E-37	122	52.46
gi_22531389_emb_CAD44625.1_	cathepsin B1 isotype 2 [Sm]	340	3	gi_18181863_emb_CAC85211.2_	cathepsin B endopeptidase [Sm]	1.00E-100	311	51.77
gi_18181863_emb_CAC85211.2_	cathepsin B endopeptidase [Sm]	347	4	gi_22531389_emb_CAD44625.1_	cathepsin B1 isotype 2 [Sm]	1.00E-100	311	51.77

Appendix D: Sequence Homology Between Sample Sets

Secretion proteins with homology to tegument proteins

(More than 50% homology over at least 100 residues with an expectation value of 1x10-5 or better)

accession	description	length	num hits	hit accession	hit description	hit significance	hsp length	% identity
AAD17299.1	thioredoxin peroxidase [Sm]	185	3	TC14049	TC2639 TC288 TC3226 TC3474 TC5109 frame +3	1.00E-110	185	99.46
AAP94734.1	ATP-diphosphohydrolase 1 [Sm]	544	4	TC11432	TC2288 TC3771 TC6362 frame +1	0	545	98.90

Vesicle proteins with homology to tegument proteins

(More than 50% homology over at least 100 residues with an expectation value of $1x10^{-5}$ or better)

accession	description	length	num hits	hit accession	hit description	hit significance	hsp length	% identity
AAD17299.1	thioredoxin peroxidase [Schistosoma mansoni]	185	3	TC14049	TC2639 TC288 TC3226 TC3474 TC5109 frame +3	1E-110	185	99.46
AAA29882.1	fimbrin	651	5	TC7585	TC2210 TC341 TC3368 TC6464 frame +2	0	651	99.08
TC10839	thioredoxin peroxidase	219	5	TC14049	TC2639 TC288 TC3226 TC3474 TC5109 frame +3	3E-63	179	59.78
P12812	Major egg antigen (p40)	354	4	TC7485	TC5509 frame +1	1E-111	353	54.11

Secretion proteins with homology to vesicle proteins

(More than 50% homology over at least 100 residues with an expectation value of 1x10^{-5} or better)

Accession	description	length	num hits	hit accession	hit description	hit significance	hsp length	% identity
P48501	Triosephosphate isomerase (TIM)	253	7	P48501	Triosephosphate isomerase (TIM)	1.00E-148	253	100.00
TC7546	ORF 25..776 frame +1	250	3	TC7546	ORF 25..776 frame +1	1E-148	250	100.00
TC7454	TC1083 TC1563 TC2625 TC526 TC3281 TC4409 TC4966 TC5646 frame is 3	572	3	TC7454	TC1083 TC1563 TC2625 TC526 TC3281 TC4409 TC4966 TC5646 frame is 3	0	572	100
TC7336	ORF 106..1460 frame +1	443	7	TC7336	ORF 106..1460 frame +1	0	443	100
TC16844	ORF 109..1133 frame +1	341	7	TC16844	ORF 109..1133 frame +1	0	341	100
TC16812	ORF 74..522 frame +2	149	15	TC16812	ORF 74..522 frame +2	1E-83	149	100
TC16735	ORF 98..1095 frame +2	332	6	TC16735	ORF 98..1095 frame +2	0	332	100
TC14578	ORF 95..405 frame +2	103	13	TC14578	ORF 95..405 frame +2	2E-56	103	100
TC13658	ORF 30..439 frame +3	136	8	TC13658	ORF 30..439 frame +3	7E-74	136	100
TC13606	ORF 24..391 frame +3	122	3	TC13606	ORF 24..391 frame +3	2E-65	122	100
TC13604	ORF 280..1832 frame +1	517	4	TC13604	ORF 280..1832 frame +1	0	517	100
TC13591	TC1613 TC1618 TC219 TC707 TC3313 TC4136 TC4992 TC6430 frame is 1	847	6	TC13591	TC1613 TC1618 TC219 TC707 TC3313 TC4136 TC4992 TC6430 frame is 1	0	847	100
TC11413	ORF 38..486 frame +2	149	6	TC11413	ORF 38..486 frame +2	1E-86	149	100
P16641	ATP:guanidino kinase SMC74 (ATP:guanidino phosphotransferase)	675	3	P16641	ATP:guanidino kinase SMC74 (ATP:guanidino phosphotransferase)	0	675	100
Q26565|	Peptidyl-prolyl cis-trans isomerase (PPIase) (Rotamase) (Cyclophilin) (Cyclosporin A-binding protein) (p17.7) (Smp17.7)	161	5	Q26565	Peptidyl-prolyl cis-trans isomerase (PPIase) (Rotamase) (Cyclophilin) (Cyclosporin A-binding protein) (p17.7) (Smp17.7)	1E-96	161	100
P41759	Phosphoglycerate kinase	416	5	P41759	Phosphoglycerate kinase	0	416	100

Accession	description	length	num hits	hit accession	hit description	hit significance	hsp length	% identity
P09792	Glutathione S-transferase 28 kDa (GST 28) (SM28 antigen) (Protective 28 kDa antigen) (GST class-mu)	211	8	P09792	Glutathione S-transferase 28 kDa (GST 28) (SM28 antigen) (Protective 28 kDa antigen) (GST class-mu)	1E-123	211	100
P20287	Glyceraldehyde-3-phosphate dehydrogenase (GAPDH) (Major larval surface antigen) (P-37)	338	6	P20287	Glyceraldehyde-3-phosphate dehydrogenase (GAPDH) (Major larval surface antigen) (P-37)	0	338	100
Q27877	Enolase (2-phosphoglycerate dehydratase) (2-phospho-D-glycerate hydro-lyase)	434	4	Q27877	Enolase (2-phosphoglycerate dehydratase) (2-phospho-D-glycerate hydro-lyase)	0	434	100
CAA69721	elongation factor 1-alpha [Sm]	465	8	CAA69721	elongation factor 1-alpha [Sm]	0	465	100
CAA28976	70000 mol wt antigen/hsp70 homologue (619 AA) [Sm]	619	4	CAA28976	70000 mol wt antigen/hsp70 homologue (619 AA) [Sm]	0	619	100
P53442	Fructose-bisphosphate aldolase	363	5	P53442	Fructose-bisphosphate aldolase	0	363	100
AAM43941	elastase 2a [Sm]	263	6	AAM43941	elastase 2a [Sm]	1E-153	263	100
AAL79841	thioredoxin [Sm]	106	11	AAL79841	thioredoxin [Sm]	3E-58	106	100
AAM69406	heat shock protein HSP60 [Sm]	549	13	AAM69406	heat shock protein HSP60 [Sm]	0	549	100
AAD24794	phosphoenolpyruvate carboxykinase [Sm]	626	2	AAD24794	phosphoenolpyruvate carboxykinase [Sm]	0	626	100
AAD26122	SPO-1 protein [Sm]	117	8	AAD26122	SPO-1 protein [Sm]	4E-65	117	100
AAR26703	actin-binding/filamin-like protein [Sm]	984	4	AAR26703	actin-binding/filamin-like protein [Sm]	0	984	100
AAD17299	thioredoxin peroxidase [Sm]	185	4	AAD17299	thioredoxin peroxidase [Sm]	1E-110	185	100
AAC79131	tegumental protein Sm 20.8 [Sm]	181	10	AAC79131	tegumental protein Sm 20.8 [Sm]	1E-104	181	100
AAC46967	elastase	274	4	AAC46967	elastase	1E-162	274	100
AAC46966	actin	376	5	AAC46966	actin	0	376	100
AAB86571	unknown [Sm] (serpin)	256	2	AAB86571	unknown [Sm] (serpin)	1E-148	256	100
AAB47536	calponin homolog [Sm]	361	4	AAB47536	calponin homolog [Sm]	0	361	100
AAB41442	putative cytosol aminopeptidase [Sm]	520	6	AAB41442	putative cytosol aminopeptidase [Sm]	0	520	100

Accession	description	length	num hits	hit accession	hit description	hit significance	hsp length	% identity
AAB21173	glutathione S-transferase GST [Sm Peptide 218 aa]	218	3	AAB21173	glutathione S-transferase GST [Sm Peptide 218 aa]	1E-132	218	100
AAA29921	calcium binding protein [Sm]	154	10	AAA29921	calcium binding protein [Sm]	3E-89	154	100
A45630	vaccine-dominant antigen Sm21.7 - fluke (Sm)	184	10	A45630	vaccine-dominant antigen Sm21.7 - fluke (Sm)	1E-107	184	100
A45529	heat shock protein 86 - fluke (Sm) (fragment)	442	8	A45529	heat shock protein 86 - fluke (Sm) (fragment)	0	442	100
A28942	pancreatic elastase (EC 3.4.21.36) precursor - fluke (Sm)	264	8	A28942	pancreatic elastase (EC 3.4.21.36) precursor - fluke (Sm)	1E-156	264	100
Q26540	14-3-3 protein homolog 1	252	7	Q26540	14-3-3 protein homolog 1	1E-143	252	100
AAC46967	elastase	274	4	A28942	pancreatic elastase (EC 3.4.21.36) precursor - fluke (Sm)	1E-144	275	92.36364
A28942	pancreatic elastase (EC 3.4.21.36) precursor - fluke (Sm)	264	8	AAC46967	elastase	1E-144	275	92.36364
AAM43941	elastase 2a [Sm]	263	6	A28942	pancreatic elastase (EC 3.4.21.36) precursor - fluke (Sm)	1E-124	259	82.23938
A28942	pancreatic elastase (EC 3.4.21.36) precursor - fluke (Sm)	264	8	AAM43941	elastase 2a [Sm]	1E-124	259	82.23938
AAM43941	elastase 2a [Sm]	263	6	AAC46967	elastase	1E-124	274	79.92701
AAC46967	elastase	274	4	AAM43941	elastase 2a [Sm]	1E-124	274	79.92701
AAD17299	thioredoxin peroxidase [Sm]	185	4	TC10839	ORF 71..729 frame +2	7E-63	179	59.77654
Q26551	Peptidyl-prolyl cis-trans isomerase B precursor (PPIase) (Rotamase) (Cyclophilin B) (S-cyclophilin)	213	5	Q26565	Peptidyl-prolyl cis-trans isomerase (PPIase) (Rotamase) (Cyclophilin) (Cyclosporin A-binding protein) (p17.7) (Smp17.7)	3E-45	164	56.70732
Q26540	14-3-3 protein homolog 1	252	7	AAF21436	14-3-3 epsilon [Sm]	1E-69	239	54.81172

Appendix E: ProP Propeptide prediction results

Only showing results for proteins predicted to contain one or more propeptides

```
A28942 pancreatic elastase
MSNRWRFVVVVTLFTYCLTFERVSTWLIRSGEPVQHPAEFPFIAFLTTERTMCTGSLVSTRAVLTAGHCVCSPLPVIRVS      80
FLTLRNGDQQGIHHQPSGVKVAPGYMPSCMSARQRRPIAQTLSGFDIAIVMLAQMVNLQSGIRVISLPQPSDIPPPGTGV     160
FIVGYGRDDNDRDPSRKNGGILKKGRATIMECRHATNGNPICVKAGQNFGQLPAPGDSGGPLLPSLQGPVLGVVSHGVTL     240
PNLPDIIVEYASVARMLDFVRSNI                                                             320
sssssssssssssssssssssssss...................................P...................      80
...................................P............................................     160
................................................................................     240
........................                                                             320

Signal peptide cleavage site predicted:        between pos. 25 and 26: VST-WL
Propeptide cleavage sites predicted:   Arg(R)/Lys(K): 2

Name           Pos    Context     Score  Pred
____________________________v_________________
A28942           4    ---MSNR|WR  0.041    .
A28942           6    -MSNRWR|FV  0.028    .
A28942          22    YCLTFER|VS  0.035    .
A28942          29    VSTWLIR|SG  0.113    .
A28942          50    AFLTTER|TM  0.056    .
A28942          61    GSLVSTR|AV  0.853 *ProP*
A28942          78    SPLPVIR|VS  0.047    .
A28942          85    VSFLTLR|NG  0.048    .
A28942         100    HQPSGVK|VA  0.088    .
A28942         113    PSCMSAR|QR  0.131    .
A28942         115    CMSARQR|RP  0.040    .
A28942         116    MSARQRR|PI  0.743 *ProP*
A28942         143    NLQSGIR|VI  0.062    .
A28942         167    FIVGYGR|DD  0.038    .
A28942         172    GRDDNDR|DP  0.055    .
A28942         176    NDRDPSR|KN  0.070    .
A28942         177    DRDPSRK|NG  0.137    .
```

```
A28942         183    KNGGILK|KG  0.101    .
A28942         184    NGGILKK|GR  0.156    .
A28942         186    GILKKGR|AT  0.062    .
A28942         193    ATIMECR|HA  0.031    .
A28942         204    GNPICVK|AG  0.055    .
A28942         255    EYASVAR|ML  0.063    .
A28942         261    RMLDFVR|SN  0.087    .
____________________________^_________________
```

AAA19024 calreticulin

```
MLSILLTLLLSKYALGHEVWFSETFPNESIENWVQSTYNAEKQGEFKVEAGKSPVDPIEDLGLKTTQDARFYGIARKISE      80
PFSNRGKTILLQFTVKFDKTVSCGGAYIKLLGSDIDPKKFHGESPYKIMFGPDICGMATKKVHVIFNYKGKNHLIKKEIP     160
CKDDLKTHLYTLIVNPNNKYEVLVDNADPNDKKPDDWVDEQFIDDPDDKKPDNWDQPKTIPDMDAKKPDDWDDAMDGEWE     240
RPQKDNPEYKGEWTPRRIDNPKYKGEWKPVQIDNPEYKHDPELYVLNDIGYVGFDLWQVDSGSIFDNILITDSPDFAKEE     320
GERLWRKRYDAEVAKEQSSAKDDKEEAEETKERKELPDDAKASDEPSGDHDEL                                400
ssssssssssssssss................................................................      80
................................................................................     160
................................................................................     240
................................................................................     320
.......P.............................................                                400

Signal peptide cleavage site predicted:        between pos. 16 and 17: ALG-HE
Propeptide cleavage sites predicted:   Arg(R)/Lys(K): 1

Name           Pos    Context     Score  Pred
____________________________v_________________
AAA19024        12    LTLLLSK|YA  0.075    .
AAA19024        42    STYNAEK|QG  0.063    .
AAA19024        47    EKQGEFK|VE  0.067    .
AAA19024        52    FKVEAGK|SP  0.072    .
AAA19024        64    IEDLGLK|TT  0.088    .
AAA19024        70    KTTQDAR|FY  0.035    .
AAA19024        76    RFYGIAR|KI  0.032    .
AAA19024        77    FYGIARK|IS  0.076    .
AAA19024        85    SEPFSNR|GK  0.053    .
AAA19024        87    PFSNRGK|TI  0.101    .
```

```
AAA19024        96    LLQFTVK|FD  0.072    .
AAA19024        99    FTVKFDK|TV  0.059    .
AAA19024       109    CGGAYIK|LL  0.065    .
AAA19024       118    GSDIDPK|KF  0.078    .
AAA19024       119    SDIDPKK|FH  0.102    .
AAA19024       127    HGESPYK|IM  0.097    .
AAA19024       140    ICGMATK|KV  0.064    .
AAA19024       141    CGMATKK|VH  0.071    .
AAA19024       149    HVIFNYK|GK  0.083    .
AAA19024       151    IFNYKGK|NH  0.074    .
AAA19024       156    GKNHLIK|KE  0.090    .
AAA19024       157    KNHLIKK|EI  0.280    .
AAA19024       162    KKEIPCK|DD  0.063    .
AAA19024       166    PCKDDLK|TH  0.113    .
AAA19024       179    IVNPNNK|YE  0.270    .
AAA19024       192    NADPNDK|KP  0.061    .
AAA19024       193    ADPNDKK|PD  0.068    .
AAA19024       209    IDDPDDK|KP  0.056    .
AAA19024       210    DDPDDKK|PD  0.067    .
AAA19024       218    DNWDQPK|TI  0.173    .
AAA19024       226    IPDMDAK|KP  0.071    .
AAA19024       227    PDMDAKK|PD  0.077    .
AAA19024       241    MDGEWER|PQ  0.045    .
AAA19024       244    EWERPQK|DN  0.109    .
AAA19024       250    KDNPEYK|GE  0.070    .
AAA19024       256    KGEWTPR|RI  0.056    .
AAA19024       257    GEWTPRR|ID  0.058    .
AAA19024       262    RRIDNPK|YK  0.083    .
AAA19024       264    IDNPKYK|GE  0.081    .
AAA19024       268    KYKGEWK|PV  0.051    .
AAA19024       278    IDNPEYK|HD  0.060    .
AAA19024       318    DSPDFAK|EE  0.106    .
AAA19024       323    AKEEGER|LW  0.048    .
AAA19024       326    EGERLWR|KR  0.025    .
AAA19024       327    GERLWRK|RY  0.064    .
AAA19024       328    ERLWRKR|YD  0.530 *ProP*
```

```
AAA19024       335    YDAEVAK|EQ  0.076    .
AAA19024       341    KEQSSAK|DD  0.076    .
AAA19024       344    SSAKDDK|EE  0.086    .
AAA19024       351    EEAEETK|ER  0.112    .
AAA19024       353    AEETKER|KE  0.071    .
AAA19024       354    EETKERK|EL  0.076    .
AAA19024       361    ELPDDAK|AS  0.099    .
____________________________^_________________
```

AAA29921 calcium binding protein

```
MAFKIDDFTIQEDQVKIAKDVFKRFDKRGQEKISTTDLGPAFRALNLTVKPDTLKEWADQVDDDATGFIDFNGFLICYGK      80
KLQEDQDERDLRDAFRVLDKNKRGEIDVEDLRWILKGLGDDLTEEEIDDMIRDTDTGSGFVDFDEFYKLMTSE            160
...........................P....................................................      80
......................P............................P.....................            160

Signal peptide cleavage site predicted:       none
Propeptide cleavage sites predicted:   Arg(R)/Lys(K): 3

Name           Pos    Context     Score  Pred
____________________________v_________________
AAA29921         4    ---MAFK|ID  0.056    .
AAA29921        16    IQEDQVK|IA  0.170    .
AAA29921        19    DQVKIAK|DV  0.064    .
AAA29921        23    IAKDVFK|RF  0.070    .
AAA29921        24    AKDVFKR|FD  0.204    .
AAA29921        27    VFKRFDK|RG  0.058    .
AAA29921        28    FKRFDKR|GQ  0.617 *ProP*
AAA29921        32    DKRGQEK|IS  0.071    .
AAA29921        43    DLGPAFR|AL  0.051    .
AAA29921        50    ALNLTVK|PD  0.067    .
AAA29921        55    VKPDTLK|EW  0.122    .
AAA29921        80    FLICYGK|KL  0.051    .
AAA29921        81    LICYGKK|LQ  0.086    .
AAA29921        89    QEDQDER|DL  0.070    .
AAA29921        92    QDERDLR|DA  0.101    .
AAA29921        96    DLRDAFR|VL  0.070    .
```

```
AAA29921       100    AFRVLDK|NK  0.062    .
AAA29921       102    RVLDKNK|RG  0.074    .
AAA29921       103    VLDKNKR|GE  0.580 *ProP*
AAA29921       112    IDVEDLR|WI  0.044    .
AAA29921       116    DLRWILK|GL  0.107    .
AAA29921       132    EIDDMIR|DT  0.634 *ProP*
AAA29921       149    DFDEFYK|LM  0.066    .
____________________________^_________________
```

AAC46967 elastase

```
MSNRWRFLVVTLFTYCLTFERVSTWLIRSGEPVQHRTEFPFIAFLTTERTMCTGSLVSTRAVLTAGHCVCSPLPVIRVLC      80
LFQVSFLTLRNGDQQGIHHQPSGVKVAPGYMPSCMSARRGRPIAQTLSGFDIAIVMLAQMVNLQSGITVISLPQASDIPT     160
PGTGVFIVGYGRDDNDRDPSRKNGGILKKGELVVGRATIMECRHATNGNPICVKAGQNFGQLPAPGDSGGPLLPSPQGPV     240
LGVVSHGVTLPNLPDIIVEYASVARMLDFVRSNI                                                   320
ssssssssssssssssssssssss...................................P....................      80
................................................................................     160
................................................................................     240
..................................                                                   320

Signal peptide cleavage site predicted:        between pos. 24 and 25: VST-WL
Propeptide cleavage sites predicted:   Arg(R)/Lys(K): 1

Name           Pos    Context     Score  Pred
____________________________v_________________
AAC46967         4    ---MSNR|WR  0.036    .
AAC46967         6    -MSNRWR|FL  0.033    .
AAC46967        21    YCLTFER|VS  0.035    .
AAC46967        28    VSTWLIR|SG  0.113    .
AAC46967        36    GEPVQHR|TE  0.151    .
AAC46967        49    AFLTTER|TM  0.056    .
AAC46967        60    GSLVSTR|AV  0.853 *ProP*
AAC46967        77    SPLPVIR|VL  0.136    .
AAC46967        90    VSFLTLR|NG  0.048    .
AAC46967       105    HQPSGVK|VA  0.088    .
AAC46967       118    PSCMSAR|RG  0.054    .
AAC46967       119    SCMSARR|GR  0.045    .
```

```
AAC46967       121    MSARRGR|PI  0.392    .
AAC46967       172    FIVGYGR|DD  0.038    .
AAC46967       177    GRDDNDR|DP  0.055    .
AAC46967       181    NDRDPSR|KN  0.070    .
AAC46967       182    DRDPSRK|NG  0.137    .
AAC46967       188    KNGGILK|KG  0.083    .
AAC46967       189    NGGILKK|GE  0.117    .
AAC46967       196    GELVVGR|AT  0.093    .
AAC46967       203    ATIMECR|HA  0.031    .
AAC46967       214    GNPICVK|AG  0.055    .
AAC46967       265    EYASVAR|ML  0.063    .
AAC46967       271    RMLDFVR|SN  0.087    .
____________________________^_________________
```

AAD17299 thioredoxin peroxidase 2

```
MVLLPNRPAPEFKGQAVINGEFKEICLKDYRGKYVVLFFYPADFTFVCPTEIIAFSDQVEEFNSRNCQVIACSTDSQYSH       80
LAWDNLDRKSGGLGHMKIPLLADRKQEISKAYGVFDEEDGNAFRGLFIIDPNGILRQITINDKPVGRSVDETLRLLDAFQ      160
FVEKHGEVCPVNWKRGQHGIKVNQK                                                             240
................................................................................       80
..................................................................P.............      160
.........................                                                             240

Signal peptide cleavage site predicted:        none
Propeptide cleavage sites predicted:   Arg(R)/Lys(K): 1

Name           Pos    Context     Score  Pred
____________________________v_________________
AAD17299         7    MVLLPNR|PA  0.126    .
AAD17299        13    RPAPEFK|GQ  0.066    .
AAD17299        23    VINGEFK|EI  0.089    .
AAD17299        28    FKEICLK|DY  0.062    .
AAD17299        31    ICLKDYR|GK  0.034    .
AAD17299        33    LKDYRGK|YV  0.087    .
AAD17299        65    VEEFNSR|NC  0.101    .
AAD17299        88    AWDNLDR|KS  0.023    .
AAD17299        89    WDNLDRK|SG  0.155    .
```

```
AAD17299        97    GGLGHMK|IP  0.070    .
AAD17299       104    IPLLADR|KQ  0.037    .
AAD17299       105    PLLADRK|QE  0.069    .
AAD17299       110    RKQEISK|AY  0.080    .
AAD17299       124    EDGNAFR|GL  0.048    .
AAD17299       136    DPNGILR|QI  0.043    .
AAD17299       143    QITINDK|PV  0.051    .
AAD17299       147    NDKPVGR|SV  0.634 *ProP*
AAD17299       154    SVDETLR|LL  0.135    .
AAD17299       164    AFQFVEK|HG  0.075    .
AAD17299       174    VCPVNWK|RG  0.062    .
AAD17299       175    CPVNWKR|GQ  0.373    .
AAD17299       181    RGQHGIK|VN  0.066    .
AAD17299       185    GIKVNQK|--  0.080    .
____________________________^_________________
```

AAF82607 AUT1

```
MAGVDHAYQYSVEVKSRFSLFLDDTLNSEDPDILLSKLQSKRGEKTKKDKPHLQQQHVAPTKADTITKSEVKLDTATPKA      80
GSRVSKTPNSTEPPPVPPEDVQITSAKGTDEPISTFXRGRGSGRGTPRGMRVGRGQGPRIAPTEAPQDSVSDLNAPRGSS     160
FEPRGRGRGRGRGMFGRGRGMPFNSNRDFENQDGPDRQGPRQYGRRDGNWNSQDVDGLIMPESGDSEQVVRFADDRNEVE     240
DQPEHATAENEEGVVVGTETPVEEEPKSYTLEGYKAMRQSSKPAVLLNNKGLRKANDGKDVFANMVAHRKLQEVSEDVYE     320
VEERKTSLTYVSFLCIYTRFHLGASVDRY                                                       400
................................................................................      80
.....................................................P..........................     160
................................................................................     240
................................................................................     320
.............................                                                       400

Signal peptide cleavage site predicted:        none
Propeptide cleavage sites predicted:   Arg(R)/Lys(K): 1

Name           Pos    Context     Score  Pred
____________________________v_________________
AAF82607        15    QYSVEVK|SR  0.070    .
AAF82607        17    SVEVKSR|FS  0.142    .
AAF82607        37    PDILLSK|LQ  0.084    .
```

```
AAF82607        41    LSKLQSK|RG  0.088    .
AAF82607        42    SKLQSKR|GE  0.185    .
AAF82607        45    QSKRGEK|TK  0.062    .
AAF82607        47    KRGEKTK|KD  0.076    .
AAF82607        48    RGEKTKK|DK  0.176    .
AAF82607        50    EKTKKDK|PH  0.060    .
AAF82607        62    QHVAPTK|AD  0.085    .
AAF82607        68    KADTITK|SE  0.155    .
AAF82607        72    ITKSEVK|LD  0.098    .
AAF82607        79    LDTATPK|AG  0.073    .
AAF82607        83    TPKAGSR|VS  0.048    .
AAF82607        86    AGSRVSK|TP  0.211    .
AAF82607       107    VQITSAK|GT  0.081    .
AAF82607       118    PISTFXR|GR  0.100    .
AAF82607       120    STFXRGR|GS  0.037    .
AAF82607       124    RGRGSGR|GT  0.062    .
AAF82607       128    SGRGTPR|GM  0.054    .
AAF82607       131    GTPRGMR|VG  0.083    .
AAF82607       134    RGMRVGR|GQ  0.722 *ProP*
AAF82607       139    GRGQGPR|IA  0.166    .
AAF82607       157    SDLNAPR|GS  0.094    .
AAF82607       164    GSSFEPR|GR  0.269    .
AAF82607       166    SFEPRGR|GR  0.236    .
AAF82607       168    EPRGRGR|GR  0.046    .
AAF82607       170    RGRGRGR|GR  0.056    .
AAF82607       172    RGRGRGR|GM  0.038    .
AAF82607       177    GRGMFGR|GR  0.113    .
AAF82607       179    GMFGRGR|GM  0.038    .
AAF82607       187    MPFNSNR|DF  0.048    .
AAF82607       197    NQDGPDR|QG  0.041    .
AAF82607       201    PDRQGPR|QY  0.038    .
AAF82607       205    GPRQYGR|RD  0.032    .
AAF82607       206    PRQYGRR|DG  0.080    .
AAF82607       231    DSEQVVR|FA  0.046    .
AAF82607       236    VRFADDR|NE  0.037    .
AAF82607       267    PVEEEPK|SY  0.074    .
```

```
AAF82607       275    YTLEGYK|AM  0.064    .
AAF82607       278    EGYKAMR|QS  0.038    .
AAF82607       282    AMRQSSK|PA  0.066    .
AAF82607       290    AVLLNNK|GL  0.093    .
AAF82607       293    LNNKGLR|KA  0.043    .
AAF82607       294    NNKGLRK|AN  0.159    .
AAF82607       299    RKANDGK|DV  0.066    .
AAF82607       309    ANMVAHR|KL  0.038    .
AAF82607       310    NMVAHRK|LQ  0.069    .
AAF82607       324    VYEVEER|KT  0.034    .
AAF82607       325    YEVEERK|TS  0.076    .
AAF82607       339    FLCIYTR|FH  0.033    .
AAF82607       348    LGASVDR|Y-  0.033    .
____________________________^_________________
```

AAM43941 elastase 2a

```
MLNGRTFLMVTLFTYCLTFERVSTWLVRKGEPVQDRTEFPYIAFVRTERTMCTGSLVSTRAVLTAGHCVCSPMPVVQVSF       80
LTLRNGDQQGIHHQPSGVKVAPEYMPSCTASRQRRRIRQTLSGFDIATVMLAQMVNLQSGIRVISLPQASDIPTPGTDVF      160
IVGYGRDDNDRDPSRRAGGILKKGRATVMECKHSTTGNPICVQAAYVFGQITAPGDSGGPLLRSPQGPVLGVVSHGVTLS      240
NRLDVLVEYASVARMLGFVSSNI                                                               320
ssssssssssssssssssssssss...................................P....................       80
................................................................................      160
...............P................................................................      240
.......................                                                               320

Signal peptide cleavage site predicted:        between pos. 24 and 25: VST-WL
Propeptide cleavage sites predicted:   Arg(R)/Lys(K): 2

Name           Pos    Context     Score  Pred
____________________________v_________________
AAM43941         5    --MLNGR|TF  0.050    .
AAM43941        21    YCLTFER|VS  0.037    .
AAM43941        28    VSTWLVR|KG  0.040    .
AAM43941        29    STWLVRK|GE  0.138    .
AAM43941        36    GEPVQDR|TE  0.053    .
AAM43941        46    PYIAFVR|TE  0.042    .
```

```
AAM43941        49    AFVRTER|TM  0.113    .
AAM43941        60    GSLVSTR|AV  0.853 *ProP*
AAM43941        84    VSFLTLR|NG  0.048    .
AAM43941        99    HQPSGVK|VA  0.104    .
AAM43941       112    PSCTASR|QR  0.059    .
AAM43941       114    CTASRQR|RR  0.045    .
AAM43941       115    TASRQRR|RI  0.222    .
AAM43941       116    ASRQRRR|IR  0.207    .
AAM43941       118    RQRRRIR|QT  0.130    .
AAM43941       142    NLQSGIR|VI  0.062    .
AAM43941       166    FIVGYGR|DD  0.038    .
AAM43941       171    GRDDNDR|DP  0.052    .
AAM43941       175    NDRDPSR|RA  0.076    .
AAM43941       176    DRDPSRR|AG  0.849 *ProP*
AAM43941       182    RAGGILK|KG  0.102    .
AAM43941       183    AGGILKK|GR  0.271    .
AAM43941       185    GILKKGR|AT  0.098    .
AAM43941       192    ATVMECK|HS  0.056    .
AAM43941       223    SGGPLLR|SP  0.166    .
AAM43941       242    GVTLSNR|LD  0.046    .
AAM43941       254    EYASVAR|ML  0.081    .
____________________________^_________________
```

TC14049 thioredoxin peroxidase 1a

```
MVLLPNRPAPEFKGQAVINGEFKEICLKDYRGKYVVLFFYPSDFTFVCPTEIIAFSDQVEEFNSRNCQVIACSTDSQYSH      80
LAWDNLDRKSGGLGHMKIPLLADRKQEISKAYGVFDEEDGNAFRGLFIIDPNGILRQITINDKPVGRSVDETLRLLDAFQ     160
FVEKHGEVCPVNWKRGQHGIKVNQK                                                           240
................................................................................      80
..................................................................P.............     160
.........................                                                           240

Signal peptide cleavage site predicted:        none
Propeptide cleavage sites predicted:   Arg(R)/Lys(K): 1

Name           Pos    Context     Score  Pred
____________________________v_________________
```

```
TC14049          7    MVLLPNR|PA  0.126    .
TC14049         13    RPAPEFK|GQ  0.066    .
TC14049         23    VINGEFK|EI  0.089    .
TC14049         28    FKEICLK|DY  0.062    .
TC14049         31    ICLKDYR|GK  0.034    .
TC14049         33    LKDYRGK|YV  0.087    .
TC14049         65    VEEFNSR|NC  0.101    .
TC14049         88    AWDNLDR|KS  0.023    .
TC14049         89    WDNLDRK|SG  0.155    .
TC14049         97    GGLGHMK|IP  0.070    .
TC14049        104    IPLLADR|KQ  0.037    .
TC14049        105    PLLADRK|QE  0.069    .
TC14049        110    RKQEISK|AY  0.080    .
TC14049        124    EDGNAFR|GL  0.048    .
TC14049        136    DPNGILR|QI  0.043    .
TC14049        143    QITINDK|PV  0.051    .
TC14049        147    NDKPVGR|SV  0.634 *ProP*
TC14049        154    SVDETLR|LL  0.135    .
TC14049        164    AFQFVEK|HG  0.075    .
TC14049        174    VCPVNWK|RG  0.062    .
TC14049        175    CPVNWKR|GQ  0.373    .
TC14049        181    RGQHGIK|VN  0.066    .
TC14049        185    GIKVNQK|--  0.080    .
_____________________________^_________________
```

TC7546 phosphoglycerate mutase

```
MAPYRIVFIRHGESVYNEENRFCGWHDADLSGQGITEAKQAGQLLRQNHFTFDIAYTSVLKRAIKTLNFVLDELDLNWIP      80
VTKTWRLNERMYGALQGLNKSETAAKHGEEQVKIWRRAYDIPPPPVDISDPRFPGNEPKYALLDSSCIPRTECLKDTVQR     160
VLPFWFDTISASIKRREQVLIVAHGNSLRALIKYLDNTSDSDIVELNIPTGIPLVYELDANLKPTKHYYLADEATVAAAI     240
ARVANQGKKK                                                                           320
.............................................................P..................      80
................................................................................     160
................................................................................     240
..........                                                                           320
Signal peptide cleavage site predicted:        none
Propeptide cleavage sites predicted:   Arg(R)/Lys(K): 1
```

```
Name               Pos    Context     Score  Pred
____________________________v_________________
TC7546               5    --MAPYR|IV  0.033     .
TC7546              10    YRIVFIR|HG  0.030     .
TC7546              21    VYNEENR|FC  0.146     .
TC7546              39    QGITEAK|QA  0.067     .
TC7546              46    QAGQLLR|QN  0.086     .
TC7546              61    AYTSVLK|RA  0.067     .
TC7546              62    YTSVLKR|AI  0.753  *ProP*
TC7546              65    VLKRAIK|TL  0.119     .
TC7546              83    NWIPVTK|TW  0.056     .
TC7546              86    PVTKTWR|LN  0.030     .
TC7546              90    TWRLNER|MY  0.033     .
TC7546             100    ALQGLNK|SE  0.073     .
TC7546             106    KSETAAK|HG  0.061     .
TC7546             113    HGEEQVK|IW  0.091     .
TC7546             116    EQVKIWR|RA  0.042     .
TC7546             117    QVKIWRR|AY  0.079     .
TC7546             132    VDISDPR|FP  0.079     .
TC7546             139    FPGNEPK|YA  0.099     .
TC7546             150    DSSCIPR|TE  0.058     .
TC7546             155    PRTECLK|DT  0.084     .
TC7546             160    LKDTVQR|VL  0.104     .
TC7546             174    TISASIK|RR  0.076     .
TC7546             175    ISASIKR|RE  0.063     .
TC7546             176    SASIKRR|EQ  0.365     .
TC7546             189    AHGNSLR|AL  0.051     .
TC7546             193    SLRALIK|YL  0.067     .
TC7546             223    ELDANLK|PT  0.061     .
TC7546             226    ANLKPTK|HY  0.069     .
TC7546             242    VAAAIAR|VA  0.068     .
TC7546             248    RVANQGK|KK  0.068     .
TC7546             249    VANQGKK|K-  0.060     .
TC7546             250    ANQGKKK|--  0.075     .
                                  ^
____________________________ _________________
```

Appendix F: Gene Ontology Annotation

VESICLES			
Biological Process			
10	12.35%	GO:0006457	protein folding
8	9.88%	GO:0006096	glycolysis
6	7.41%	GO:0044267	cellular protein metabolism
4	4.94%	GO:0006508	proteolysis
3	3.70%	GO:0007001	chromosome organization and biogenesis [sensu Eukaryota]
3	3.70%	GO:0006118	electron transport
3	3.70%	GO:0008152	metabolism
3	3.70%	GO:0006334	nucleosome assembly
2	2.47%	GO:0006754	ATP biosynthesis
2	2.47%	GO:0015986	ATP synthesis coupled proton transport
2	2.47%	GO:0005975	carbohydrate metabolism
2	2.47%	GO:0006812	cation transport
2	2.47%	GO:0007018	microtubule-based movement
2	2.47%	GO:0006412	protein biosynthesis
2	2.47%	GO:0051258	protein polymerization
2	2.47%	GO:0006100	tricarboxylic acid cycle intermediate metabolism
1	1.23%	GO:0031032	actomyosin structure organization and biogenesis
1	1.23%	GO:0019642	anaerobic glycolysis
1	1.23%	GO:0006527	arginine catabolism
1	1.23%	GO:0007155	cell adhesion
1	1.23%	GO:0001539	ciliary or flagellar motility
1	1.23%	GO:0006241	CTP biosynthesis
1	1.23%	GO:0006094	gluconeogenesis
1	1.23%	GO:0006183	GTP biosynthesis
1	1.23%	GO:0006092	main pathways of carbohydrate metabolism

TEGUMENT			
Biological Process			
3	6.98%	GO:0006810	transport
2	4.65%	GO:0006464	protein modification
1	2.33%	GO:0006887	exocytosis
1	2.33%	GO:0048278	vesicle docking
1	2.33%	GO:0006412	protein biosynthesis
1	2.33%	GO:0009117	nucleotide metabolism
1	2.33%	GO:0008643	carbohydrate transport
1	2.33%	GO:0008152	metabolism
1	2.33%	GO:0006605	protein targeting
Cellular Component			
4	9.30%	GO:0016021	integral to membrane
3	6.98%	GO:0016020	membrane
2	4.65%	GO:0005622	intracellular
1	2.33%	GO:0005576	extracellular region
1	2.33%	GO:0005737	cytoplasm
1	2.33%	GO:0005840	ribosome
1	2.33%	GO:0005795	Golgi stack
Molecular Function			
2	4.65%	GO:0005509	calcium ion binding
2	4.65%	GO:0005215	transporter activity
2	4.65%	GO:0016787	hydrolase activity
1	2.33%	GO:0005544	calcium-dependent phospholipid binding
1	2.33%	GO:0004040	amidase activity
1	2.33%	GO:0004719	protein-L-isoaspartate [D-aspartate] O-methyltransferase activity
1	2.33%	GO:0003735	structural constituent of ribosome

1	1.23%	GO:0006108	malate metabolism
1	1.23%	GO:0007017	microtubule-based process
1	1.23%	GO:0045978	negative regulation of nucleoside metabolism
1	1.23%	GO:0006807	nitrogen compound metabolism
1	1.23%	GO:0006468	protein amino acid phosphorylation
1	1.23%	GO:0019538	protein metabolism
1	1.23%	GO:0009209	pyrimidine ribonucleoside triphosphate biosynthesis
1	1.23%	GO:0006810	transport
1	1.23%	GO:0006099	tricarboxylic acid cycle
1	1.23%	GO:0006228	UTP biosynthesis
Cellular Component			
5	6.17%	GO:0005737	cytoplasm
3	3.70%	GO:0016020	membrane
3	3.70%	GO:0000786	nucleosome
3	3.70%	GO:0005634	nucleus
2	2.47%	GO:0005622	intracellular
2	2.47%	GO:0005874	microtubule
2	2.47%	GO:0016459	myosin
2	2.47%	GO:0043234	protein complex
2	2.47%	GO:0016469	proton-transporting two-sector ATPase complex
1	1.23%	GO:0005884	actin filament
1	1.23%	GO:0005783	endoplasmic reticulum
1	1.23%	GO:0009288	flagellum
1	1.23%	GO:0045255	hydrogen-translocating F-type ATPase complex
1	1.23%	GO:0016021	integral to membrane
1	1.23%	GO:0005875	microtubule associated complex
1	1.23%	GO:0005743	mitochondrial inner membrane
1	1.23%	GO:0005739	mitochondrion
1	1.23%	GO:0000015	phosphopyruvate hydratase complex
1	1.23%	GO:0005840	ribosome

1	2.33%	GO:0005554	molecular function unknown
1	2.33%	GO:0005515	protein binding
1	2.33%	GO:0005351	sugar porter activity
1	2.33%	GO:0003824	catalytic activity
1	2.33%	GO:0003676	nucleic acid binding
1	2.33%	GO:0004386	helicase activity
1	2.33%	GO:0005524	ATP binding
CONTROLS			
Biological Process			
11	7.10%	GO:0006355	regulation of transcription
9	5.81%	GO:0006468	protein amino acid phosphorylation
8	5.16%	GO:0008152	metabolism
7	4.52%	GO:0006812	cation transport
7	4.52%	GO:0006810	transport
6	3.87%	GO:0006811	ion transport
6	3.87%	GO:0006508	proteolysis
5	3.23%	GO:0006816	calcium ion transport
5	3.23%	GO:0006118	electron transport
5	3.23%	GO:0007186	G-protein coupled receptor protein signaling pathway
5	3.23%	GO:0006457	protein folding
4	2.58%	GO:0007242	intracellular signaling cascade
4	2.58%	GO:0007017	microtubule-based process
4	2.58%	GO:0006412	protein biosynthesis
4	2.58%	GO:0006278	RNA-dependent DNA replication
3	1.94%	GO:0006310	DNA recombination
3	1.94%	GO:0045449	regulation of transcription
3	1.94%	GO:0007264	small GTPase mediated signal transduction
2	1.29%	GO:0006486	protein amino acid glycosylation
2	1.29%	GO:0006413	translational initiation
2	1.29%	GO:0016192	vesicle-mediated transport

Molecular Function			
16	19.75%	GO:0005524	ATP binding
9	11.11%	GO:0051082	unfolded protein binding
7	8.64%	GO:0005509	calcium ion binding
7	8.64%	GO:0005515	protein binding
4	4.94%	GO:0005525	GTP binding
3	3.70%	GO:0003824	catalytic activity
3	3.70%	GO:0003677	DNA binding
3	3.70%	GO:0016820	hydrolase activity
3	3.70%	GO:0003774	motor activity
3	3.70%	GO:0000166	nucleotide binding
3	3.70%	GO:0016491	oxidoreductase activity
3	3.70%	GO:0004252	serine-type endopeptidase activity
2	2.47%	GO:0003779	actin binding
2	2.47%	GO:0015662	ATPase activity
2	2.47%	GO:0005489	electron transporter activity
2	2.47%	GO:0003924	GTPase activity
2	2.47%	GO:0046933	hydrogen-transporting ATP synthase activity
2	2.47%	GO:0046961	hydrogen-transporting ATPase activity
2	2.47%	GO:0016868	intramolecular transferase activity
2	2.47%	GO:0004459	L-lactate dehydrogenase activity
2	2.47%	GO:0003676	nucleic acid binding
2	2.47%	GO:0017111	nucleoside-triphosphatase activity
2	2.47%	GO:0019904	protein domain specific binding
2	2.47%	GO:0005198	structural molecule activity
1	1.23%	GO:0004177	aminopeptidase activity
1	1.23%	GO:0004053	arginase activity
1	1.23%	GO:0005488	binding
1	1.23%	GO:0004108	citrate [Si]-synthase activity
1	1.23%	GO:0004857	enzyme inhibitor activity

1	0.65%	GO:0031032	actomyosin structure organization and biogenesis
1	0.65%	GO:0006865	amino acid transport
1	0.65%	GO:0006527	arginine catabolism
1	0.65%	GO:0009072	aromatic amino acid family metabolism
1	0.65%	GO:0015986	ATP synthesis coupled proton transport
1	0.65%	GO:0005975	carbohydrate metabolism
1	0.65%	GO:0007155	cell adhesion
1	0.65%	GO:0045454	cell redox homeostasis
1	0.65%	GO:0006633	fatty acid biosynthesis
1	0.65%	GO:0006094	gluconeogenesis
1	0.65%	GO:0006424	glutamyl-tRNA aminoacylation
1	0.65%	GO:0046168	glycerol-3-phosphate catabolism
1	0.65%	GO:0006072	glycerol-3-phosphate metabolism
1	0.65%	GO:0006629	lipid metabolism
1	0.65%	GO:0015672	monovalent inorganic cation transport
1	0.65%	GO:0045978	negative regulation of nucleoside metabolism
1	0.65%	GO:0006836	neurotransmitter transport
1	0.65%	GO:0006139	nucleobase
1	0.65%	GO:0009405	pathogenesis
1	0.65%	GO:0006518	peptide metabolism
1	0.65%	GO:0006813	potassium ion transport
1	0.65%	GO:0006464	protein modification
1	0.65%	GO:0015031	protein transport
1	0.65%	GO:0015992	proton transport
1	0.65%	GO:0006979	response to oxidative stress
1	0.65%	GO:0042427	serotonin biosynthesis
1	0.65%	GO:0007165	signal transduction
1	0.65%	GO:0006414	translational elongation
1	0.65%	GO:0006418	tRNA aminoacylation for protein translation
1	0.65%	GO:0000160	two-component signal transduction system [phosphorelay]

1	1.23%	GO:0004332	fructose-bisphosphate aldolase activity
1	1.23%	GO:0004356	glutamate-ammonia ligase activity
1	1.23%	GO:0004365	glyceraldehyde-3-phosphate dehydrogenase [phosphorylating] activity
1	1.23%	GO:0004386	helicase activity
1	1.23%	GO:0008553	hydrogen-exporting ATPase activity
1	1.23%	GO:0016787	hydrolase activity
1	1.23%	GO:0016301	kinase activity
1	1.23%	GO:0004178	leucyl aminopeptidase activity
1	1.23%	GO:0030060	L-malate dehydrogenase activity
1	1.23%	GO:0000287	magnesium ion binding
1	1.23%	GO:0016615	malate dehydrogenase activity
1	1.23%	GO:0030145	manganese ion binding
1	1.23%	GO:0003777	microtubule motor activity
1	1.23%	GO:0005554	molecular function unknown
1	1.23%	GO:0051287	NAD binding
1	1.23%	GO:0004550	nucleoside diphosphate kinase activity
1	1.23%	GO:0004611	phosphoenolpyruvate carboxykinase activity
1	1.23%	GO:0004618	phosphoglycerate kinase activity
1	1.23%	GO:0004634	phosphopyruvate hydratase activity
1	1.23%	GO:0004645	phosphorylase activity
1	1.23%	GO:0004672	protein kinase activity
1	1.23%	GO:0030170	pyridoxal phosphate binding
1	1.23%	GO:0004743	pyruvate kinase activity
1	1.23%	GO:0004867	serine-type endopeptidase inhibitor activity
1	1.23%	GO:0008236	serine-type peptidase activity
1	1.23%	GO:0005200	structural constituent of cytoskeleton
1	1.23%	GO:0003735	structural constituent of ribosome
1	1.23%	GO:0046912	transferase activity
1	1.23%	GO:0016772	transferase activity
1	1.23%	GO:0004802	transketolase activity

1	0.65%	GO:0007601	visual perception
Cellular Component			
18	11.61%	GO:0016020	membrane
16	10.32%	GO:0016021	integral to membrane
12	7.74%	GO:0005634	nucleus
5	3.23%	GO:0005622	intracellular
4	2.58%	GO:0005875	microtubule associated complex
3	1.94%	GO:0005737	cytoplasm
2	1.29%	GO:0045211	postsynaptic membrane
2	1.29%	GO:0005840	ribosome
2	1.29%	GO:0005891	voltage-gated calcium channel complex
1	0.65%	GO:0000785	chromatin
1	0.65%	GO:0005783	endoplasmic reticulum
1	0.65%	GO:0005850	eukaryotic translation initiation factor 2 complex
1	0.65%	GO:0005576	extracellular region
1	0.65%	GO:0005615	extracellular space
1	0.65%	GO:0009331	glycerol-3-phosphate dehydrogenase complex
1	0.65%	GO:0005834	heterotrimeric G-protein complex
1	0.65%	GO:0005887	integral to plasma membrane
1	0.65%	GO:0005739	mitochondrion
1	0.65%	GO:0016469	proton-transporting two-sector ATPase complex
1	0.65%	GO:0008076	voltage-gated potassium channel complex
Molecular Function			
18	11.61%	GO:0005524	ATP binding
12	7.74%	GO:0003677	DNA binding
9	5.81%	GO:0003824	catalytic activity
9	5.81%	GO:0004672	protein kinase activity
8	5.16%	GO:0005509	calcium ion binding
7	4.52%	GO:0003700	transcription factor activity
6	3.87%	GO:0005525	GTP binding

1	1.23%	GO:0005215	transporter activity
1	1.23%	GO:0004807	triose-phosphate isomerase activity

SECRETIONS

Biological Process

10	18.87%	GO:0006096	glycolysis
5	9.43%	GO:0006508	proteolysis
4	7.55%	GO:0006457	protein folding
3	5.66%	GO:0006334	nucleosome assembly
3	5.66%	GO:0007001	chromosome organization and biogenesis
3	5.66%	GO:0006100	tricarboxylic acid cycle intermediate metabolism
2	3.77%	GO:0008152	metabolism
2	3.77%	GO:0006108	malate metabolism
1	1.89%	GO:0001539	ciliary or flagellar motility
1	1.89%	GO:0031032	actomyosin structure organization and biogenesis
1	1.89%	GO:0006801	superoxide metabolism
1	1.89%	GO:0006094	gluconeogenesis
1	1.89%	GO:0006810	transport
1	1.89%	GO:0006118	electron transport
1	1.89%	GO:0044267	cellular protein metabolism
1	1.89%	GO:0006412	protein biosynthesis
1	1.89%	GO:0006826	iron ion transport
1	1.89%	GO:0006879	iron ion homeostasis
1	1.89%	GO:0007017	microtubule-based process
1	1.89%	GO:0006183	GTP biosynthesis
1	1.89%	GO:0006228	UTP biosynthesis
1	1.89%	GO:0006241	CTP biosynthesis
1	1.89%	GO:0009209	pyrimidine ribonucleoside triphosphate biosynthesis
1	1.89%	GO:0005975	carbohydrate metabolism
1	1.89%	GO:0006754	ATP biosynthesis
1	1.89%	GO:0015986	ATP synthesis coupled proton transport

6	3.87%	GO:0005216	ion channel activity
5	3.23%	GO:0003676	nucleic acid binding
5	3.23%	GO:0004713	protein-tyrosine kinase activity
5	3.23%	GO:0003723	RNA binding
4	2.58%	GO:0015662	ATPase activity
4	2.58%	GO:0005489	electron transporter activity
4	2.58%	GO:0016820	hydrolase activity
4	2.58%	GO:0003777	microtubule motor activity
4	2.58%	GO:0016491	oxidoreductase activity
4	2.58%	GO:0004674	protein serine/threonine kinase activity
4	2.58%	GO:0003964	RNA-directed DNA polymerase activity
3	1.94%	GO:0005388	calcium-transporting ATPase activity
3	1.94%	GO:0005261	cation channel activity
3	1.94%	GO:0001584	rhodopsin-like receptor activity
3	1.94%	GO:0003743	translation initiation factor activity
3	1.94%	GO:0051082	unfolded protein binding
3	1.94%	GO:0008270	zinc ion binding
2	1.29%	GO:0003779	actin binding
2	1.29%	GO:0004190	aspartic-type endopeptidase activity
2	1.29%	GO:0004197	cysteine-type endopeptidase activity
2	1.29%	GO:0008234	cysteine-type peptidase activity
2	1.29%	GO:0005230	extracellular ligand-gated ion channel activity
2	1.29%	GO:0008417	fucosyltransferase activity
2	1.29%	GO:0016787	hydrolase activity
2	1.29%	GO:0004879	ligand-dependent nuclear receptor activity
2	1.29%	GO:0004497	monooxygenase activity
2	1.29%	GO:0030594	neurotransmitter receptor activity
2	1.29%	GO:0004889	nicotinic acetylcholine-activated cation-selective channel activity
2	1.29%	GO:0005515	protein binding
2	1.29%	GO:0019904	protein domain specific binding

1	1.89%	GO:0019642	anaerobic glycolysis
1	1.89%	GO:0007018	microtubule-based movement
1	1.89%	GO:0051258	protein polymerization
Cellular Component			
3	5.66%	GO:0005622	intracellular
3	5.66%	GO:0005737	cytoplasm
3	5.66%	GO:0000786	nucleosome
3	5.66%	GO:0005634	nucleus
1	1.89%	GO:0009288	flagellum
1	1.89%	GO:0005884	actin filament
1	1.89%	GO:0005945	6-phosphofructokinase complex
1	1.89%	GO:0000015	phosphopyruvate hydratase complex
1	1.89%	GO:0005875	microtubule associated complex
1	1.89%	GO:0016021	integral to membrane
1	1.89%	GO:0045255	hydrogen-translocating F-type ATPase complex
1	1.89%	GO:0016469	proton-transporting two-sector ATPase complex
1	1.89%	GO:0005874	microtubule
1	1.89%	GO:0043234	protein complex
Molecular Function			
5	9.43%	GO:0005509	calcium ion binding
4	7.55%	GO:0016491	oxidoreductase activity
4	7.55%	GO:0005524	ATP binding
3	5.66%	GO:0004252	serine-type endopeptidase activity
3	5.66%	GO:0005525	GTP binding
3	5.66%	GO:0003677	DNA binding
3	5.66%	GO:0004459	L-lactate dehydrogenase activity
2	3.77%	GO:0051082	unfolded protein binding
2	3.77%	GO:0003779	actin binding
2	3.77%	GO:0005515	protein binding
2	3.77%	GO:0005488	binding
2	1.29%	GO:0004252	serine-type endopeptidase activity
2	1.29%	GO:0004871	signal transducer activity
2	1.29%	GO:0003707	steroid hormone receptor activity
2	1.29%	GO:0003735	structural constituent of ribosome
2	1.29%	GO:0005245	voltage-gated calcium channel activity
1	0.65%	GO:0004040	amidase activity
1	0.65%	GO:0015359	amino acid permease activity
1	0.65%	GO:0005279	amino acid-polyamine transporter activity
1	0.65%	GO:0004053	arginase activity
1	0.65%	GO:0042626	ATPase activity
1	0.65%	GO:0016887	ATPase activity
1	0.65%	GO:0005488	binding
1	0.65%	GO:0005544	calcium-dependent phospholipid binding
1	0.65%	GO:0005386	carrier activity
1	0.65%	GO:0004104	cholinesterase activity
1	0.65%	GO:0005507	copper ion binding
1	0.65%	GO:0004869	cysteine protease inhibitor activity
1	0.65%	GO:0015036	disulfide oxidoreductase activity
1	0.65%	GO:0004866	endopeptidase inhibitor activity
1	0.65%	GO:0004857	enzyme inhibitor activity
1	0.65%	GO:0005006	epidermal growth factor receptor activity
1	0.65%	GO:0050660	FAD binding
1	0.65%	GO:0015018	galactosylgalactosylxylosylprotein 3-beta-glucuronosyltransferase activity
1	0.65%	GO:0004818	glutamate-tRNA ligase activity
1	0.65%	GO:0004819	glutamine-tRNA ligase activity
1	0.65%	GO:0004602	glutathione peroxidase activity
1	0.65%	GO:0004367	glycerol-3-phosphate dehydrogenase [NAD+] activity
1	0.65%	GO:0019001	guanyl nucleotide binding
1	0.65%	GO:0031072	heat shock protein binding
1	0.65%	GO:0046933	hydrogen-transporting ATP synthase activity

2	3.77%	GO:0030060	L-malate dehydrogenase activity
2	3.77%	GO:0016615	malate dehydrogenase activity
1	1.89%	GO:0004198	calpain activity
1	1.89%	GO:0004197	cysteine-type endopeptidase activity
1	1.89%	GO:0004177	aminopeptidase activity
1	1.89%	GO:0004867	serine-type endopeptidase inhibitor activity
1	1.89%	GO:0004785	copper, zinc superoxide dismutase activity
1	1.89%	GO:0046872	metal ion binding
1	1.89%	GO:0003774	motor activity
1	1.89%	GO:0005200	structural constituent of cytoskeleton
1	1.89%	GO:0004611	phosphoenolpyruvate carboxykinase activity
1	1.89%	GO:0008289	lipid binding
1	1.89%	GO:0005489	electron transporter activity
1	1.89%	GO:0008236	serine-type peptidase activity
1	1.89%	GO:0016787	hydrolase activity
1	1.89%	GO:0004869	cysteine protease inhibitor activity
1	1.89%	GO:0004866	endopeptidase inhibitor activity
1	1.89%	GO:0016301	kinase activity
1	1.89%	GO:0016772	transferase activity
1	1.89%	GO:0004365	glyceraldehyde-3-phosphate dehydrogenase, phosphorylating activity
1	1.89%	GO:0051287	NAD binding
1	1.89%	GO:0008199	ferric iron binding
1	1.89%	GO:0004618	phosphoglycerate kinase activity
1	1.89%	GO:0004807	triose-phosphate isomerase activity
1	1.89%	GO:0004332	fructose-bisphosphate aldolase activity
1	1.89%	GO:0019904	protein domain specific binding
1	1.89%	GO:0003872	6-phosphofructokinase activity
1	1.89%	GO:0004634	phosphopyruvate hydratase activity
1	1.89%	GO:0003777	microtubule motor activity

1	0.65%	GO:0046961	hydrogen-transporting ATPase activity
1	0.65%	GO:0005506	iron ion binding
1	0.65%	GO:0016301	kinase activity
1	0.65%	GO:0008289	lipid binding
1	0.65%	GO:0015077	monovalent inorganic cation transporter activity
1	0.65%	GO:0008080	N-acetyltransferase activity
1	0.65%	GO:0051287	NAD binding
1	0.65%	GO:0004983	neuropeptide Y receptor activity
1	0.65%	GO:0005328	neurotransmitter-sodium symporter activity
1	0.65%	GO:0017111	nucleoside-triphosphatase activity
1	0.65%	GO:0000166	nucleotide binding
1	0.65%	GO:0016654	oxidoreductase activity
1	0.65%	GO:0016616	oxidoreductase activity
1	0.65%	GO:0016614	oxidoreductase activity
1	0.65%	GO:0004504	peptidylglycine monooxygenase activity
1	0.65%	GO:0004611	phosphoenolpyruvate carboxykinase activity
1	0.65%	GO:0016791	phosphoric monoester hydrolase activity
1	0.65%	GO:0005267	potassium channel activity
1	0.65%	GO:0004731	purine-nucleoside phosphorylase activity
1	0.65%	GO:0004872	receptor activity
1	0.65%	GO:0008236	serine-type peptidase activity
1	0.65%	GO:0005198	structural molecule activity
1	0.65%	GO:0005529	sugar binding
1	0.65%	GO:0016772	transferase activity
1	0.65%	GO:0016763	transferase activity, transferring pentosyl groups
1	0.65%	GO:0003746	translation elongation factor activity
1	0.65%	GO:0004675	transmembrane receptor protein serine/threonine kinase activity
1	0.65%	GO:0004714	transmembrane receptor protein tyrosine kinase activity
1	0.65%	GO:0005215	transporter activity

1	1.89%	GO:0004550	nucleoside diphosphate kinase activity
1	1.89%	GO:0000287	magnesium ion binding
1	1.89%	GO:0004645	phosphorylase activity
1	1.89%	GO:0030170	pyridoxal phosphate binding
1	1.89%	GO:0008553	hydrogen-exporting ATPase activity
1	1.89%	GO:0046933	hydrogen-transporting ATP synthase activity
1	1.89%	GO:0046961	hydrogen-transporting ATPase activity
1	1.89%	GO:0000166	nucleotide binding
1	1.89%	GO:0017111	nucleoside-triphosphatase activity
1	1.89%	GO:0005198	structural molecule activity
1	1.89%	GO:0003924	GTPase activity
1	1.89%	GO:0004743	pyruvate kinase activity
1	1.89%	GO:0016868	intramolecular transferase activity

1	0.65%	GO:0004812	tRNA ligase activity
1	0.65%	GO:0004510	tryptophan 5-monooxygenase activity
1	0.65%	GO:0005249	voltage-gated potassium channel activity

Appendix G: Protein Domain Annotation

VESICLES					CONTROLS			
6	7.41%	IPR002423	Chaperonin Cpn60/TCP-1		9	20.93%	IPR000719	Protein kinase
6	7.41%	IPR008950	GroEL-like chaperone, ATPase		8	18.60%	IPR011009	Protein kinase-like
6	7.41%	IPR011992	EF-Hand type		7	16.28%	IPR012335	Thioredoxin fold
6	7.41%	IPR012335	Thioredoxin fold		6	13.95%	IPR011992	EF-Hand type
5	6.17%	IPR001844	Chaperonin Cpn60		5	11.63%	IPR005834	Haloacid dehalogenase-like hydrolase
4	4.94%	IPR002194	Chaperonin TCP-1		5	11.63%	IPR002048	Calcium-binding EF-hand
4	4.94%	IPR002048	Calcium-binding EF-hand		5	11.63%	IPR001245	Tyrosine protein kinase
3	3.70%	IPR007125	Histone core		4	9.30%	IPR004014	Cation transporting ATPase, N-terminal
3	3.70%	IPR009072	Histone-fold		4	9.30%	IPR008250	E1-E2 ATPase-associated region
3	3.70%	IPR001314	Peptidase S1A, chymotrypsin		4	9.30%	IPR001757	ATPase, E1-E2 type
3	3.70%	IPR001254	Peptidase S1 and S6, chymotrypsin/Hap		4	9.30%	IPR006068	Cation transporting ATPase, C-terminal
3	3.70%	IPR009003	Peptidase, trypsin-like serine and cysteine		4	9.30%	IPR006663	Thioredoxin domain 2
2	2.47%	IPR008280	Tubulin/FtsZ, C-terminal		4	9.30%	IPR006662	Thioredoxin-related
2	2.47%	IPR000217	Tubulin		4	9.30%	IPR001372	Dynein light chain, type 1
2	2.47%	IPR003008	Tubulin/FtsZ, GTPase		4	9.30%	IPR008266	Tyrosine protein kinase, active site
2	2.47%	IPR001557	L-lactate/malate dehydrogenase		4	9.30%	IPR000477	RNA-directed DNA polymerase Reverse transcriptase
2	2.47%	IPR001236	Lactate/malate dehydrogenase		3	6.98%	IPR005225	Small GTP-binding protein domain
2	2.47%	IPR000194	H+-transporting two-sector ATPase, alpha/beta subunit, central region		3	6.98%	IPR001806	Ras GTPase
2	2.47%	IPR003593	AAA ATPase		3	6.98%	IPR007087	Zinc finger, C2H2-type
2	2.47%	IPR004100	H+-transporting two-sector ATPase, alpha/beta subunit, N-terminal		3	6.98%	IPR008271	Serine/threonine protein kinase, active site
2	2.47%	IPR000793	H+-transporting two-sector ATPase, alpha/beta subunit, C-terminal		3	6.98%	IPR005821	Ion transport
2	2.47%	IPR000866	Alkyl hydroperoxide reductase/ Thiol specific antioxidant/ Mal allergen		3	6.98%	IPR005820	Cation channel, non-ligand gated
2	2.47%	IPR000308	14-3-3 protein		3	6.98%	IPR012336	Thioredoxin-like fold
2	2.47%	IPR000533	Tropomyosin		3	6.98%	IPR000276	Rhodopsin-like GPCR superfamily

2	2.47%	IPR010987	Glutathione S-transferase, C-terminal-like		3	6.98%	IPR000980	SH2 motif
2	2.47%	IPR004046	Glutathione S-transferase, C-terminal		3	6.98%	IPR001611	Leucine-rich repeat
2	2.47%	IPR004045	Glutathione S-transferase, N-terminal		3	6.98%	IPR001356	Homeobox
2	2.47%	IPR002928	Myosin tail		3	6.98%	IPR009057	Homeodomain-like
2	2.47%	IPR006662	Thioredoxin-related		3	6.98%	IPR012287	Homeodomain-related
2	2.47%	IPR000886	Endoplasmic reticulum targeting sequence		3	6.98%	IPR002290	Serine/threonine protein kinase
2	2.47%	IPR005834	Haloacid dehalogenase-like hydrolase		3	6.98%	IPR001584	Integrase, catalytic region
2	2.47%	IPR004014	Cation transporting ATPase, N-terminal		2	4.65%	IPR001388	Synaptobrevin
2	2.47%	IPR001757	ATPase, E1-E2 type		2	4.65%	IPR001452	Src homology-3
2	2.47%	IPR006068	Cation transporting ATPase, C-terminal		2	4.65%	IPR001680	WD-40 repeat
2	2.47%	IPR008250	E1-E2 ATPase-associated region		2	4.65%	IPR013019	MAD homology, MH1
2	2.47%	IPR009079	Four-helical cytokine		2	4.65%	IPR008984	SMAD/FHA
2	2.47%	IPR001404	Heat shock protein Hsp90		2	4.65%	IPR003619	Dwarfin protein, A
2	2.47%	IPR001715	Calponin-like actin-binding		2	4.65%	IPR001132	Dwarfin protein
1	1.23%	IPR013078	Phosphoglycerate mutase		2	4.65%	IPR009356	NADH dehydrogenase subunit 4L
1	1.23%	IPR005952	Phosphoglycerate mutase 1		2	4.65%	IPR001682	Ca2+/Na+ channel, pore region
1	1.23%	IPR001650	Helicase, C-terminal		2	4.65%	IPR002111	Cation not K+ channel, TM region
1	1.23%	IPR011545	DEAD/DEAH box helicase, N-terminal		2	4.65%	IPR002077	Ca2+ channel, alpha subunit
1	1.23%	IPR001697	Pyruvate kinase		2	4.65%	IPR013027	FAD-dependent pyridine nucleotide-disulphide oxidoreductase
1	1.23%	IPR001978	Troponin		2	4.65%	IPR000308	14-3-3 protein
1	1.23%	IPR000158	Cell division protein FtsZ		2	4.65%	IPR013128	Peptidase C1A, papain
1	1.23%	IPR002453	Beta tubulin		2	4.65%	IPR000668	Peptidase C1A, papain C-terminal
1	1.23%	IPR012718	T-complex protein 1, epsilon subunit		2	4.65%	IPR000169	Peptidase, cysteine peptidase active site
1	1.23%	IPR007648	Mitochondrial ATPase inhibitor, IATP		2	4.65%	IPR012599	Peptidase C1, propeptide
1	1.23%	IPR012722	T-complex protein 1, zeta subunit		2	4.65%	IPR001715	Calponin-like actin-binding
1	1.23%	IPR001993	Mitochondrial substrate carrier		2	4.65%	IPR005746	Thioredoxin
1	1.23%	IPR002113	Adenine nucleotide translocator 1		2	4.65%	IPR009003	Peptidase, trypsin-like serine and cysteine
1	1.23%	IPR002067	Mitochondrial carrier protein		2	4.65%	IPR001314	Peptidase S1A, chymotrypsin
1	1.23%	IPR001252	Malate dehydrogenase, active site		2	4.65%	IPR001254	Peptidase S1 and S6, chymotrypsin/Hap
1	1.23%	IPR010097	Malate dehydrogenase, NAD-dependent, eukaryotes and gamma proteobacteria		2	4.65%	IPR000379	Esterase/lipase/thioesterase

1	1.23%	IPR001125	Recoverin		2	4.65%	IPR008139	Saposin B
1	1.23%	IPR011304	L-lactate dehydrogenase		2	4.65%	IPR011001	Saposin-like
1	1.23%	IPR005475	Transketolase, central region		2	4.65%	IPR000301	CD9/CD37/CD63 antigen
1	1.23%	IPR009014	Transketolase, C-terminal-like		2	4.65%	IPR008952	Tetraspanin
1	1.23%	IPR005476	Transketolase, C-terminal		2	4.65%	IPR000555	Mov34/MPN/PAD-1
1	1.23%	IPR005474	Transketolase, N-terminal		2	4.65%	IPR008991	Translation protein SH3-like
1	1.23%	IPR005478	Bacterial transketolase		2	4.65%	IPR001884	Eukaryotic initiation factor 5A hypusine eIF-5A
1	1.23%	IPR001951	Histone H4		2	4.65%	IPR005824	KOW
1	1.23%	IPR003953	Fumarate reductase/succinate dehydrogenase flavoprotein, N-terminal		2	4.65%	IPR013032	EGF-like region
1	1.23%	IPR004112	Fumarate reductase/succinate dehydrogenase flavoprotein, C-terminal		2	4.65%	IPR001503	Glycosyl transferase, family 10
1	1.23%	IPR012719	T-complex protein 1, gamma subunit		2	4.65%	IPR001827	Homeobox protein, antennapedia type
1	1.23%	IPR000164	Histone H3		2	4.65%	IPR005782	Calcium ATPase
1	1.23%	IPR012716	T-complex protein 1, beta subunit		2	4.65%	IPR006202	Neurotransmitter-gated ion-channel ligand-binding
1	1.23%	IPR000558	Histone H2B		2	4.65%	IPR006029	Neurotransmitter-gated ion-channel transmembrane region
1	1.23%	IPR005722	ATP synthase F1, beta subunit		2	4.65%	IPR006201	Neurotransmitter-gated ion-channel
1	1.23%	IPR000811	Glycosyl transferase, family 35		2	4.65%	IPR002394	Nicotinic acetylcholine receptor
1	1.23%	IPR011833	Glycogen/starch/alpha-glucan phosphorylase		2	4.65%	IPR008946	Steroid nuclear receptor, ligand-binding
1	1.23%	IPR001564	Nucleoside diphosphate kinase		2	4.65%	IPR001628	Nuclear hormone receptor, DNA-binding
1	1.23%	IPR012005	Nucleoside-diphosphate kinase		2	4.65%	IPR000536	Nuclear hormone receptor, ligand-binding
1	1.23%	IPR012336	Thioredoxin-like fold		2	4.65%	IPR001723	Steroid hormone receptor
1	1.23%	IPR000911	Ribosomal protein L11		2	4.65%	IPR000324	Vitamin D receptor
1	1.23%	IPR001107	Band 7 protein		2	4.65%	IPR008042	Retrotransposon, Pao
1	1.23%	IPR000163	Prohibitin		2	4.65%	IPR001995	Peptidase A2A, retrovirus, catalytic
1	1.23%	IPR002020	Citrate synthase		2	4.65%	IPR009007	Peptidase aspartic, catalytic
1	1.23%	IPR010109	Citrate synthase, eukaryotic		2	4.65%	IPR000886	Endoplasmic reticulum targeting sequence
1	1.23%	IPR005841	Phosphoglucomutase/phosphomannomutase		2	4.65%	IPR002579	Methionine sulfoxide reductase B
1	1.23%	IPR005843	Phosphoglucomutase/phosphomannomutase C terminal		2	4.65%	IPR011057	Mss4-like

1	1.23%	IPR005845	Phosphoglucomutase/phosphomannomutase alpha/beta/alpha domain II		1	2.33%	IPR005027	Glycosyl transferase, family 43
1	1.23%	IPR005846	Phosphoglucomutase/phosphomannomutase alpha/beta/alpha domain III		1	2.33%	IPR006383	HAD-superfamily hydrolase subfamily IB, PSPase-like
1	1.23%	IPR005844	Phosphoglucomutase/phosphomannomutase alpha/beta/alpha domain I		1	2.33%	IPR002347	Glucose/ribitol dehydrogenase
1	1.23%	IPR003960	AAA-protein subdomain		1	2.33%	IPR002198	Short-chain dehydrogenase/reductase SDR
1	1.23%	IPR003338	AAA ATPase VAT, N-terminal		1	2.33%	IPR007648	Mitochondrial ATPase inhibitor, IATP
1	1.23%	IPR009010	Aspartate decarboxylase-like fold		1	2.33%	IPR003579	Ras small GTPase, Rab type
1	1.23%	IPR003959	AAA ATPase, central region		1	2.33%	IPR002078	Sigma-54 factor, interaction region
1	1.23%	IPR005938	AAA ATPase, CDC48		1	2.33%	IPR000866	Alkyl hydroperoxide reductase/ Thiol specific antioxidant/ Mal allergen
1	1.23%	IPR005294	ATP synthase F1, alpha subunit		1	2.33%	IPR011511	Variant SH3
1	1.23%	IPR000504	RNA-binding region RNP-1 RNA recognition motif		1	2.33%	IPR012615	Trematode Eggshell Synthesis
1	1.23%	IPR012677	Nucleotide-binding, alpha-beta plait		1	2.33%	IPR001369	Purine phosphorylase, family 2
1	1.23%	IPR001372	Dynein light chain, type 1		1	2.33%	IPR011270	Purine nucleoside phosphorylase I, inosine and guanosine-specific
1	1.23%	IPR000941	Enolase		1	2.33%	IPR011268	Inosine guanosine and xanthosine phosphorylase
1	1.23%	IPR002130	Peptidyl-prolyl cis-trans isomerase, cyclophilin type		1	2.33%	IPR000782	Beta-Ig-H3/fasciclin
1	1.23%	IPR000741	Fructose-bisphosphate aldolase, class-I		1	2.33%	IPR006413	Calcium-transporting P-type ATPase, PMR1-type
1	1.23%	IPR000652	Triosephosphate isomerase		1	2.33%	IPR000695	H+ transporting ATPase, proton pump
1	1.23%	IPR001576	Phosphoglycerate kinase		1	2.33%	IPR003605	TGF beta receptor, GS motif
1	1.23%	IPR000173	Glyceraldehyde 3-phosphate dehydrogenase		1	2.33%	IPR001289	CCAAT-binding transcription factor, subunit B
1	1.23%	IPR000749	ATP:guanido phosphotransferase		1	2.33%	IPR002086	Aldehyde dehydrogenase
1	1.23%	IPR008978	HSP20-like chaperone		1	2.33%	IPR005829	Sugar transporter superfamily
1	1.23%	IPR001436	Alpha crystallin		1	2.33%	IPR005446	L-type voltage-dependent calcium channel alpha 1 subunit
1	1.23%	IPR002068	Heat shock protein Hsp20		1	2.33%	IPR000205	NAD-binding site
1	1.23%	IPR000719	Protein kinase		1	2.33%	IPR001327	Pyridine nucleotide-disulphide oxidoreductase, NAD-binding region
1	1.23%	IPR011009	Protein kinase-like		1	2.33%	IPR012999	Pyridine nucleotide-disulphide oxidoreductase, class I, active site

1	1.23%	IPR004161	Elongation factor Tu, domain 2		1	2.33%	IPR011767	Glutaredoxin active site
1	1.23%	IPR009000	Translation factor		1	2.33%	IPR011899	Glutaredoxin, eukaryotic and viruses
1	1.23%	IPR000795	Protein synthesis factor, GTP-binding		1	2.33%	IPR006338	Thioredoxin and glutathione reductase selenoprotein
1	1.23%	IPR009001	EF-Tu/eEF-1alpha/eIF2-gamma, C-terminal		1	2.33%	IPR002109	Glutaredoxin
1	1.23%	IPR004160	Elongation factor Tu, C-terminal		1	2.33%	IPR004099	Pyridine nucleotide-disulphide oxidoreductase dimerisation region
1	1.23%	IPR013126	Heat shock protein 70		1	2.33%	IPR001100	Pyridine nucleotide-disulphide oxidoreductase, class I
1	1.23%	IPR001023	Heat shock protein Hsp70		1	2.33%	IPR000815	Mercuric reductase
1	1.23%	IPR001298	Filamin/ABP280 repeat		1	2.33%	IPR001412	Aminoacyl-tRNA synthetase, class I
1	1.23%	IPR002173	Carbohydrate kinase, PfkB		1	2.33%	IPR000924	Glutamyl-tRNA synthetase, class Ic
1	1.23%	IPR005924	Arginase		1	2.33%	IPR011035	Ribosomal protein L25-like
1	1.23%	IPR006035	Arginase/agmatinase/formiminoglutamase		1	2.33%	IPR007638	Glutaminyl-tRNA synthetase, non-specific RNA-binding region part 2
1	1.23%	IPR008256	Peptidase S1B, glutamyl endopeptidase I		1	2.33%	IPR004514	Glutaminyl-tRNA synthetase
1	1.23%	IPR001983	Translationally controlled tumor protein		1	2.33%	IPR003577	Ras small GTPase, Ras type
1	1.23%	IPR011057	Mss4-like		1	2.33%	IPR000264	Serum albumin
1	1.23%	IPR006069	Cation transporting ATPase		1	2.33%	IPR001703	Alpha-fetoprotein
1	1.23%	IPR008209	Phosphoenolpyruvate carboxykinase GTP		1	2.33%	IPR006069	Cation transporting ATPase
1	1.23%	IPR000782	Beta-Ig-H3/fasciclin		1	2.33%	IPR005775	Na+/K+ ATPase, alpha subunit
1	1.23%	IPR004001	Actin		1	2.33%	IPR012340	Nucleic acid-binding, OB-fold, subgroup
1	1.23%	IPR004000	Actin/actin-like		1	2.33%	IPR003029	RNA binding S1
1	1.23%	IPR000215	Proteinase inhibitor I4, serpin		1	2.33%	IPR011488	Eukaryotic translation initiation factor 2, alpha subunit
1	1.23%	IPR000557	Calponin repeat		1	2.33%	IPR008994	Nucleic acid-binding, OB-fold
1	1.23%	IPR001997	Calponin		1	2.33%	IPR001623	Heat shock protein DnaJ, N-terminal
1	1.23%	IPR003096	SM22/calponin		1	2.33%	IPR004160	Elongation factor Tu, C-terminal
1	1.23%	IPR000819	Peptidase M17, leucyl aminopeptidase, C-terminal		1	2.33%	IPR004161	Elongation factor Tu, domain 2
1	1.23%	IPR011356	Peptidase M17, leucyl aminopeptidase		1	2.33%	IPR009001	EF-Tu/eEF-1alpha/eIF2-gamma, C-terminal
1	1.23%	IPR003299	Flagellar calcium-binding protein calflagin		1	2.33%	IPR000795	Protein synthesis factor, GTP-binding
1	1.23%	IPR001589	Actin-binding, actinin-type		1	2.33%	IPR009000	Translation factor

1	1.23%	IPR001580	Calreticulin/calnexin
1	1.23%	IPR013320	Concanavalin A-like lectin/glucanase, subgroup
1	1.23%	IPR008985	Concanavalin A-like lectin/glucanase
1	1.23%	IPR009169	Calreticulin
1	1.23%	IPR007420	Protein of unknown function DUF465
1	1.23%	IPR004009	Myosin, N-terminal, SH3-like
1	1.23%	IPR001609	Myosin head, motor region
1	1.23%	IPR001637	Glutamine synthetase class-I, adenylation site
1	1.23%	IPR002452	Alpha tubulin
SECRETIONS			
5	9.43%	IPR011992	EF-Hand type
4	7.55%	IPR012335	Thioredoxin fold
3	5.66%	IPR009003	Peptidase, trypsin-like serine and cysteine
3	5.66%	IPR001314	Peptidase S1A, chymotrypsin
3	5.66%	IPR001254	Peptidase S1 and S6, chymotrypsin/Hap
3	5.66%	IPR001557	L-lactate/malate dehydrogenase
3	5.66%	IPR001236	Lactate/malate dehydrogenase
3	5.66%	IPR009072	Histone-fold
3	5.66%	IPR007125	Histone core
3	5.66%	IPR002048	Calcium-binding EF-hand
2	3.77%	IPR002130	Peptidyl-prolyl cis-trans isomerase, cyclophilin type
2	3.77%	IPR001252	Malate dehydrogenase, active site
2	3.77%	IPR004045	Glutathione S-transferase, N-terminal
2	3.77%	IPR010987	Glutathione S-transferase, C-terminal-like
2	3.77%	IPR004046	Glutathione S-transferase, C-terminal
2	3.77%	IPR001715	Calponin-like actin-binding
1	1.89%	IPR003008	Tubulin/FtsZ, GTPase
1	1.89%	IPR008280	Tubulin/FtsZ, C-terminal
1	1.89%	IPR000217	Tubulin

1	2.33%	IPR004539	Translation elongation factor EF-1, alpha subunit
1	2.33%	IPR009068	S15/NS1, RNA-binding
1	2.33%	IPR001828	Extracellular ligand-binding receptor
1	2.33%	IPR001997	Calponin
1	2.33%	IPR000557	Calponin repeat
1	2.33%	IPR003096	SM22/calponin
1	2.33%	IPR011045	Nitrous oxide reductase, N-terminal
1	2.33%	IPR000463	Cytosolic fatty-acid binding
1	2.33%	IPR000566	Lipocalin-related protein and Bos/Can/Equ allergen
1	2.33%	IPR012674	Calycin
1	2.33%	IPR011038	Calycin-like
1	2.33%	IPR000175	Sodium:neurotransmitter symporter
1	2.33%	IPR000911	Ribosomal protein L11
1	2.33%	IPR008256	Peptidase S1B, glutamyl endopeptidase I
1	2.33%	IPR000120	Amidase
1	2.33%	IPR001179	Peptidylprolyl isomerase, FKBP-type
1	2.33%	IPR000759	Adrenodoxin reductase
1	2.33%	IPR003140	Phospholipase/Carboxylesterase
1	2.33%	IPR004843	Metallophosphoesterase
1	2.33%	IPR006186	Serine/threonine-specific protein phosphatase and bis5-nucleosyl-tetraphosphatase
1	2.33%	IPR011002	Flagellar motor switch protein FliG-like
1	2.33%	IPR006212	Furin-like repeat
1	2.33%	IPR000494	EGF receptor, L domain
1	2.33%	IPR003961	Fibronectin, type III
1	2.33%	IPR000585	Hemopexin
1	2.33%	IPR009030	Growth factor, receptor
1	2.33%	IPR001019	Guanine nucleotide binding protein G-protein, alpha subunit
1	2.33%	IPR001408	G-protein alpha subunit, group I
1	2.33%	IPR011025	G protein alpha subunit, helical insertion

1	1.89%	IPR000652	Triosephosphate isomerase
1	1.89%	IPR009000	Translation factor
1	1.89%	IPR006662	Thioredoxin-related
1	1.89%	IPR001424	Superoxide dismutase, copper/zinc binding
1	1.89%	IPR003096	SM22/calponin
1	1.89%	IPR001125	Recoverin
1	1.89%	IPR001697	Pyruvate kinase
1	1.89%	IPR000215	Proteinase inhibitor I4, serpin
1	1.89%	IPR001713	Proteinase inhibitor I25A, stefin A
1	1.89%	IPR000010	Proteinase inhibitor I25, cystatin
1	1.89%	IPR000795	Protein synthesis factor, GTP-binding
1	1.89%	IPR005952	Phosphoglycerate mutase 1
1	1.89%	IPR013078	Phosphoglycerate mutase
1	1.89%	IPR001576	Phosphoglycerate kinase
1	1.89%	IPR000023	Phosphofructokinase
1	1.89%	IPR008209	Phosphoenolpyruvate carboxykinase GTP
1	1.89%	IPR000169	Peptidase, cysteine peptidase active site
1	1.89%	IPR008256	Peptidase S1B, glutamyl endopeptidase I
1	1.89%	IPR000819	Peptidase M17, leucyl aminopeptidase, C-terminal
1	1.89%	IPR001300	Peptidase C2, calpain
1	1.89%	IPR012005	Nucleoside-diphosphate kinase
1	1.89%	IPR000407	Nucleoside phosphatase GDA1/CD39
1	1.89%	IPR001564	Nucleoside diphosphate kinase
1	1.89%	IPR010097	Malate dehydrogenase, NAD-dependent, eukaryotes and gamma proteobacteria
1	1.89%	IPR011274	Malate dehydrogenase, NAD-dependent, cytosolic
1	1.89%	IPR010945	Malate dehydrogenase, NAD or NADP
1	1.89%	IPR008267	Malate dehydrogenase
1	1.89%	IPR011304	L-lactate dehydrogenase

1	2.33%	IPR001770	G-protein, gamma subunit
1	2.33%	IPR003578	Ras small GTPase, Rho type
1	2.33%	IPR001569	Ribosomal protein L37e
1	2.33%	IPR008977	PHM/PNGase F Fold
1	2.33%	IPR000323	Copper type II, ascorbate-dependent monooxygenase
1	2.33%	IPR000720	Peptidyl-glycine alpha-amidating monooxygenase
1	2.33%	IPR000408	Regulator of chromosome condensation, RCC1
1	2.33%	IPR010987	Glutathione S-transferase, C-terminal-like
1	2.33%	IPR004045	Glutathione S-transferase, N-terminal
1	2.33%	IPR003957	Histone-like transcription factor/archaeal histone/topoisomerase
1	2.33%	IPR007124	Histone-fold/TFIID-TAF/NF-Y
1	2.33%	IPR009072	Histone-fold
1	2.33%	IPR003958	Transcription factor CBF/NF-Y/archaeal histone
1	2.33%	IPR008922	Di-copper centre-containing
1	2.33%	IPR002227	Tyrosinase
1	2.33%	IPR002049	EGF-like, laminin
1	2.33%	IPR000407	Nucleoside phosphatase GDA1/CD39
1	2.33%	IPR000182	GCN5-related N-acetyltransferase
1	2.33%	IPR009464	PCAF, N-terminal
1	2.33%	IPR001487	Bromodomain
1	2.33%	IPR002018	Carboxylesterase, type B
1	2.33%	IPR000997	Cholinesterase
1	2.33%	IPR001713	Proteinase inhibitor I25A, stefin A
1	2.33%	IPR000010	Proteinase inhibitor I25, cystatin
1	2.33%	IPR008967	p53-like transcription factor, DNA-binding
1	2.33%	IPR000611	Neuropeptide Y receptor
1	2.33%	IPR000794	Beta-ketoacyl synthase
1	2.33%	IPR002219	Protein kinase C, phorbol ester/diacylglycerol binding

1	1.89%	IPR000566	Lipocalin-related protein and Bos/Can/Equ allergen		1	2.33%	IPR008973	C2 calcium/lipid-binding region, CaLB
1	1.89%	IPR001951	Histone H4		1	2.33%	IPR000008	C2
1	1.89%	IPR000164	Histone H3		1	2.33%	IPR000961	Protein kinase, C-terminal
1	1.89%	IPR000558	Histone H2B		1	2.33%	IPR011991	Winged helix repressor DNA-binding
1	1.89%	IPR001404	Heat shock protein Hsp90		1	2.33%	IPR002341	HSF/ETS, DNA-binding
1	1.89%	IPR001023	Heat shock protein Hsp70		1	2.33%	IPR000232	Heat shock factor HSF-type, DNA-binding
1	1.89%	IPR013126	Heat shock protein 70		1	2.33%	IPR001298	Filamin/ABP280 repeat
1	1.89%	IPR004100	H+-transporting two-sector ATPase, alpha/beta subunit, N-terminal		1	2.33%	IPR002173	Carbohydrate kinase, PfkB
1	1.89%	IPR000793	H+-transporting two-sector ATPase, alpha/beta subunit, C-terminal		1	2.33%	IPR000194	H+-transporting two-sector ATPase, alpha/beta subunit, central region
1	1.89%	IPR000194	H+-transporting two-sector ATPase, alpha/beta subunit, central region		1	2.33%	IPR009071	High mobility group box
1	1.89%	IPR008950	GroEL-like chaperone, ATPase		1	2.33%	IPR000910	HMG1/2 high mobility group box
1	1.89%	IPR000811	Glycosyl transferase, family 35		1	2.33%	IPR000135	High mobility group proteins HMG1 and HMG2
1	1.89%	IPR011833	Glycogen/starch/alpha-glucan phosphorylase		1	2.33%	IPR000626	Ubiquitin
1	1.89%	IPR000173	Glyceraldehyde 3-phosphate dehydrogenase		1	2.33%	IPR002293	Amino acid/polyamine transporter I
1	1.89%	IPR002347	Glucose/ribitol dehydrogenase		1	2.33%	IPR004841	Amino acid permease-associated region
1	1.89%	IPR000741	Fructose-bisphosphate aldolase, class-I		1	2.33%	IPR004760	L-type amino acid transporter
1	1.89%	IPR009079	Four-helical cytokine		1	2.33%	IPR002110	Ankyrin
1	1.89%	IPR003299	Flagellar calcium-binding protein calflagin		1	2.33%	IPR003571	Snake toxin
1	1.89%	IPR001298	Filamin/ABP280 repeat		1	2.33%	IPR001273	Aromatic amino acid hydroxylase
1	1.89%	IPR012347	Ferritin-related		1	2.33%	IPR005963	Tyrosine 5-monooxygenase
1	1.89%	IPR009078	Ferritin/ribonucleotide reductase-like		1	2.33%	IPR006035	Arginase/agmatinase/formiminoglutamase
1	1.89%	IPR008331	Ferritin and Dps		1	2.33%	IPR005924	Arginase
1	1.89%	IPR001519	Ferritin		1	2.33%	IPR003593	AAA ATPase
1	1.89%	IPR000941	Enolase		1	2.33%	IPR003439	ABC transporter related
1	1.89%	IPR004161	Elongation factor Tu, domain 2		1	2.33%	IPR011527	ABC transporter, transmembrane region, type 1
1	1.89%	IPR004160	Elongation factor Tu, C-terminal		1	2.33%	IPR001140	ABC transporter, transmembrane region
1	1.89%	IPR009001	EF-Tu/eEF-1alpha/eIF2-gamma, C-terminal		1	2.33%	IPR005135	Endonuclease/exonuclease/phosphatase
1	1.89%	IPR001372	Dynein light chain, type 1		1	2.33%	IPR001878	Zinc finger, CCHC-type
1	1.89%	IPR000463	Cytosolic fatty-acid binding		1	2.33%	IPR001969	Peptidase aspartic, active site

1	1.89%	IPR002423	Chaperonin Cpn60/TCP-1
1	1.89%	IPR001844	Chaperonin Cpn60
1	1.89%	IPR000158	Cell division protein FtsZ
1	1.89%	IPR002173	Carbohydrate kinase, PfkB
1	1.89%	IPR011038	Calycin-like
1	1.89%	IPR012674	Calycin
1	1.89%	IPR000557	Calponin repeat
1	1.89%	IPR001997	Calponin
1	1.89%	IPR002453	Beta tubulin
1	1.89%	IPR000749	ATP:guanido phosphotransferase
1	1.89%	IPR005722	ATP synthase F1, beta subunit
1	1.89%	IPR000866	Alkyl hydroperoxide reductase/ Thiol specific antioxidant/ Mal allergen
1	1.89%	IPR001589	Actin-binding, actinin-type
1	1.89%	IPR004000	Actin/actin-like
1	1.89%	IPR004001	Actin
1	1.89%	IPR003593	AAA ATPase
1	1.89%	IPR009161	6-phosphofructokinase, eukaryotic type
1	1.89%	IPR000308	14-3-3 protein
TEGUMENT			
2	4.65%	IPR008973	C2 calcium/lipid-binding region, CaLB
2	4.65%	IPR012336	Thioredoxin-like fold
2	4.65%	IPR000301	CD9/CD37/CD63 antigen
2	4.65%	IPR008952	Tetraspanin
2	4.65%	IPR001715	Calponin-like actin-binding
1	2.33%	IPR001464	Annexin
1	2.33%	IPR000120	Amidase
1	2.33%	IPR012968	FerI

1	2.33%	IPR008209	Phosphoenolpyruvate carboxykinase GTP
1	2.33%	IPR001464	Annexin
1	2.33%	IPR001589	Actin-binding, actinin-type
1	2.33%	IPR001580	Calreticulin/calnexin
1	2.33%	IPR013320	Concanavalin A-like lectin/glucanase, subgroup
1	2.33%	IPR008985	Concanavalin A-like lectin/glucanase
1	2.33%	IPR009169	Calreticulin
1	2.33%	IPR003972	Shaker voltage-gated K+ channel
1	2.33%	IPR001622	K+ channel, pore region
1	2.33%	IPR003131	K+ channel tetramerisation
1	2.33%	IPR003968	Kv channel
1	2.33%	IPR003091	Voltage-dependent potassium channel
1	2.33%	IPR000210	BTB
1	2.33%	IPR007856	Saposin-like type B, 1
1	2.33%	IPR001429	ATP P2X receptor
1	2.33%	IPR000889	Glutathione peroxidase
1	2.33%	IPR007092	Leucine-rich repeat, SDS22
1	2.33%	IPR000749	ATP:guanido phosphotransferase
1	2.33%	IPR000959	POLO box duplicated region
1	2.33%	IPR008927	6-phosphogluconate dehydrogenase, C-terminal-like
1	2.33%	IPR011128	NAD-dependent glycerol-3-phosphate dehydrogenase, N-terminal
1	2.33%	IPR006109	NAD-dependent glycerol-3-phosphate dehydrogenase, C-terminal
1	2.33%	IPR006168	NAD-dependent glycerol-3-phosphate dehydrogenase
1	2.33%	IPR001304	C-type lectin
1	2.33%	IPR002130	Peptidyl-prolyl cis-trans isomerase, cyclophilin type
1	2.33%	IPR006628	PUR-alpha/beta/gamma, DNA/RNA-binding
1	2.33%	IPR009079	Four-helical cytokine

1	2.33%	IPR001283	Allergen V5/Tpx-1 related
1	2.33%	IPR012269	Aquaporin
1	2.33%	IPR000425	Major intrinsic protein
1	2.33%	IPR000682	Protein-L-isoaspartateD-aspartate O-methyltransferase
1	2.33%	IPR009976	Exocyst complex component Sec10
1	2.33%	IPR000407	Nucleoside phosphatase GDA1/CD39
1	2.33%	IPR000626	Ubiquitin
1	2.33%	IPR001975	Ribosomal protein L40e
1	2.33%	IPR001666	Phosphatidylinositol transfer protein
1	2.33%	IPR005024	Snf7
1	2.33%	IPR002017	Spectrin repeat
1	2.33%	IPR002048	Calcium-binding EF-hand
1	2.33%	IPR011992	EF-Hand type
1	2.33%	IPR000866	Alkyl hydroperoxide reductase/ Thiol specific antioxidant/ Mal allergen
1	2.33%	IPR012335	Thioredoxin fold
1	2.33%	IPR002591	Type I phosphodiesterase/nucleotide pyrophosphatase
1	2.33%	IPR001478	PDZ/DHR/GLGF
1	2.33%	IPR001298	Filamin/ABP280 repeat
1	2.33%	IPR005828	General substrate transporter
1	2.33%	IPR003663	Sugar transporter
1	2.33%	IPR005829	Sugar transporter superfamily
1	2.33%	IPR000873	AMP-dependent synthetase and ligase
1	2.33%	IPR011545	DEAD/DEAH box helicase, N-terminal
1	2.33%	IPR001650	Helicase, C-terminal
1	2.33%	IPR008978	HSP20-like chaperone
1	2.33%	IPR002068	Heat shock protein Hsp20
1	2.33%	IPR001436	Alpha crystallin
1	2.33%	IPR011710	Protein of unknown function DUF1606
1	2.33%	IPR001404	Heat shock protein Hsp90
1	2.33%	IPR003594	ATP-binding region, ATPase-like
1	2.33%	IPR001760	Opsin
1	2.33%	IPR008080	Parvalbumin
1	2.33%	IPR001125	Recoverin
1	2.33%	IPR006861	Hyaluronan/mRNA binding protein
1	2.33%	IPR001173	Glycosyl transferase, family 2

1	2.33%	IPR009028	AP2 clathrin adaptor, alpha and beta chain, appendage
1	2.33%	IPR001837	CAP protein

Appendix H: GO Entropy Calculations

(only showing attributes with highest information gain)

GO terms- Entropy: Vesicles vs. Controls

70 vesicle proteins vs. 118 control proteins

Total Information Gain = 9.52E-01

attribute description	attribute	yy	ny	nn	yn	E_attribute	Gain
Biological Process: glycolysis	GO:0006096	8	0	118	62	8.89E-01	6.30E-02
Biological Process: cellular protein metabolism	GO:0044267	6	0	118	64	9.06E-01	4.68E-02
Biological Process: regulation of transcription	GO:0006355	0	11	107	70	9.12E-01	4.09E-02
Cellular Component: integral to membrane	GO:0016021	1	16	102	69	9.14E-01	3.83E-02
Molecular Function: unfolded protein binding	GO:0051082	9	3	115	61	9.23E-01	2.91E-02
Molecular Function: transcription factor activity	GO:0003700	0	7	111	70	9.27E-01	2.56E-02
Molecular Function: protein binding	GO:0005515	7	2	116	63	9.28E-01	2.49E-02
Cellular Component: membrane	GO:0016020	3	18	100	67	9.29E-01	2.32E-02
Cellular Component: nucleosome	GO:0000786	3	0	118	67	9.29E-01	2.31E-02
Biological Process: nucleosome assembly	GO:0006334	3	0	118	67	9.29E-01	2.31E-02
Biological Process: chromosome organization and biogenesis [sensu Eukaryota]	GO:0007001	3	0	118	67	9.29E-01	2.31E-02
Molecular Function: motor activity	GO:0003774	3	0	118	67	9.29E-01	2.31E-02
Biological Process: protein folding	GO:0006457	10	5	113	60	9.30E-01	2.23E-02
Molecular Function: ion channel activity	GO:0005216	0	6	112	70	9.31E-01	2.19E-02
Biological Process: ion transport	GO:0006811	0	6	112	70	9.31E-01	2.19E-02
Biological Process: calcium ion transport	GO:0006816	0	5	113	70	9.34E-01	1.82E-02
Molecular Function: RNA binding	GO:0003723	0	5	113	70	9.34E-01	1.82E-02

attribute description	attribute	yy	ny	nn	yn	E_attribute	Gain
Biological Process: G-protein coupled receptor protein signaling pathway	GO:0007186	0	5	113	70	9.34E-01	1.82E-02
Molecular Function: protein-tyrosine kinase activity	GO:0004713	0	5	113	70	9.34E-01	1.82E-02

GO terms- Entropy: Secretions vs. Controls

48 secretion proteins vs. 118 control proteins.

Total Information Gain = 8.68E-01

attribute description	attribute	yy	ny	nn	yn	E_attribute	Gain
Biological Process: glycolysis	GO:0006096	10	0	118	38	7.53E-01	1.15E-01
Cellular Component: membrane	GO:0016020	0	18	100	48	8.10E-01	5.72E-02
Biological Process: regulation of transcription	GO:0006355	0	11	107	48	8.34E-01	3.40E-02
Cellular Component: nucleosome	GO:0000786	3	0	118	45	8.35E-01	3.29E-02
Biological Process: nucleosome assembly	GO:0006334	3	0	118	45	8.35E-01	3.29E-02
Biological Process: chromosome organization and biogenesis	GO:0007001	3	0	118	45	8.35E-01	3.29E-02
Molecular Function: L-lactate dehydrogenase activity	GO:0004459	3	0	118	45	8.35E-01	3.29E-02
Biological Process: tricarboxylic acid cycle intermediate metabolism	GO:0006100	3	0	118	45	8.35E-01	3.29E-02
Molecular Function: catalytic activity	GO:0003824	0	9	109	48	8.40E-01	2.76E-02
Molecular Function: protein kinase activity	GO:0004672	0	9	109	48	8.40E-01	2.76E-02
Biological Process: protein amino acid phosphorylation	GO:0006468	0	9	109	48	8.40E-01	2.76E-02
Cellular Component: integral to membrane	GO:0016021	1	16	102	47	8.40E-01	2.73E-02
Biological Process: malate metabolism	GO:0006108	2	0	118	46	8.46E-01	2.18E-02
Molecular Function: L-malate dehydrogenase activity	GO:0030060	2	0	118	46	8.46E-01	2.18E-02
Molecular Function: malate dehydrogenase activity	GO:0016615	2	0	118	46	8.46E-01	2.18E-02
Biological Process: cation transport	GO:0006812	0	7	111	48	8.46E-01	2.13E-02
Molecular Function: transcription factor activity	GO:0003700	0	7	111	48	8.46E-01	2.13E-02

GO terms- Entropy: Tegument vs. Controls

21 tegument proteins vs. 118 control proteins.

Total Information Gain = 6.43E-01

attribute description	attribute	yy	ny	nn	yn	E_attribute	Gain
Molecular Function: transporter activity	GO:0005215	2	1	117	19	5.91E-01	2.18E-02
Biological Process: protein modification	GO:0006464	2	1	117	19	5.91E-01	2.18E-02
Molecular Function: DNA binding	GO:0003677	0	12	106	21	5.91E-01	2.14E-02
Cellular Component: nucleus	GO:0005634	0	12	106	21	5.91E-01	2.14E-02
Molecular Function: protein-L-isoaspartate [D-aspartate] O-methyltransferase activity	GO:0004719	1	0	118	20	5.93E-01	1.98E-02
Biological Process: exocytosis	GO:0006887	1	0	118	20	5.93E-01	1.98E-02
Biological Process: vesicle docking	GO:0048278	1	0	118	20	5.93E-01	1.98E-02
Molecular Function: molecular function unknown	GO:0005554	1	0	118	20	5.93E-01	1.98E-02
Biological Process: nucleotide metabolism	GO:0009117	1	0	118	20	5.93E-01	1.98E-02
Molecular Function: sugar porter activity	GO:0005351	1	0	118	20	5.93E-01	1.98E-02
Biological Process: carbohydrate transport	GO:0008643	1	0	118	20	5.93E-01	1.98E-02
Molecular Function: helicase activity	GO:0004386	1	0	118	20	5.93E-01	1.98E-02
Cellular Component: Golgi stack	GO:0005795	1	0	118	20	5.93E-01	1.98E-02
Biological Process: protein targeting	GO:0006605	1	0	118	20	5.93E-01	1.98E-02
Biological Process: regulation of transcription	GO:0006355	0	11	107	21	5.93E-01	1.96E-02

GO terms- Entropy: Secretions vs. Vesicles

48 secretion proteins vs. 69 vesicle proteins.

Total Information Gain = 9.77E-01

attribute description	attribute	yy	ny	nn	yn	E_attribute	Gain
Molecular Function: ATP binding	GO:0005524	4	16	53	44	9.47E-01	2.93E-02
Cellular Component: membrane	GO:0016020	0	3	66	48	9.57E-01	1.99E-02
Molecular Function: hydrolase activity	GO:0016820	0	3	66	48	9.57E-01	1.99E-02
Molecular Function: catalytic activity	GO:0003824	0	3	66	48	9.57E-01	1.99E-02
Molecular Function: unfolded protein binding	GO:0051082	2	9	60	46	9.59E-01	1.78E-02
Biological Process: cellular protein metabolism	GO:0044267	1	6	63	47	9.61E-01	1.55E-02
Molecular Function: nucleic acid binding	GO:0003676	0	2	67	48	9.63E-01	1.32E-02
Cellular Component: myosin	GO:0016459	0	2	67	48	9.63E-01	1.32E-02
Biological Process: cation transport	GO:0006812	0	2	67	48	9.63E-01	1.32E-02
Molecular Function: ATPase activity	GO:0015662	0	2	67	48	9.63E-01	1.32E-02

GO terms- Entropy: Tegument vs. Vesicles

21 tegument proteins vs. 69 vesicle proteins.

Total Information Gain = 7.84E-01

attribute description	attribute	yy	ny	nn	yn	E_attribute	Gain
Molecular Function: ATP binding	GO:0005524	0	16	53	21	7.08E-01	7.62E-02
Cellular Component: integral to membrane	GO:0016021	4	1	68	17	7.22E-01	6.18E-02
Biological Process: protein modification	GO:0006464	2	0	69	19	7.36E-01	4.79E-02
Biological Process: protein folding	GO:0006457	0	10	59	21	7.38E-01	4.56E-02
Molecular Function: unfolded protein binding	GO:0051082	0	9	60	21	7.43E-01	4.07E-02
Biological Process: transport	GO:0006810	3	1	68	18	7.43E-01	4.05E-02
Biological Process: glycolysis	GO:0006096	0	8	61	21	7.48E-01	3.59E-02

attribute description	attribute	yy	ny	nn	yn	E_attribute	Gain
Biological Process: cellular protein metabolism	GO:0044267	0	6	63	21	7.57E-01	2.66E-02
Molecular Function: calcium-dependent phospholipid binding	GO:0005544	1	0	69	20	7.60E-01	2.36E-02
Molecular Function: amidase activity	GO:0004040	1	0	69	20	7.60E-01	2.36E-02
Cellular Component: extracellular region	GO:0005576	1	0	69	20	7.60E-01	2.36E-02
Molecular Function: protein-L-isoaspartate [D-aspartate] O-methyltransferase activity	GO:0004719	1	0	69	20	7.60E-01	2.36E-02
Biological Process: exocytosis	GO:0006887	1	0	69	20	7.60E-01	2.36E-02
Biological Process: vesicle docking	GO:0048278	1	0	69	20	7.60E-01	2.36E-02
Biological Process: nucleotide metabolism	GO:0009117	1	0	69	20	7.60E-01	2.36E-02
Molecular Function: sugar porter activity	GO:0005351	1	0	69	20	7.60E-01	2.36E-02
Biological Process: carbohydrate transport	GO:0008643	1	0	69	20	7.60E-01	2.36E-02
Cellular Component: Golgi stack	GO:0005795	1	0	69	20	7.60E-01	2.36E-02
Biological Process: protein targeting	GO:0006605	1	0	69	20	7.60E-01	2.36E-02
Molecular Function: transporter activity	GO:0005215	2	1	68	19	7.63E-01	2.12E-02
Molecular Function: hydrolase activity	GO:0016787	2	1	68	19	7.63E-01	2.12E-02

GO terms- Entropy: Tegument vs. Secretions

21 tegument proteins vs. 48 secretion proteins.

Total Information Gain = 8.87E-01

attribute description	attribute	yy	ny	nn	yn	E_attribute	Gain
Biological Process: glycolysis	GO:0006096	0	10	38	21	8.03E-01	8.34E-02
Cellular Component: membrane	GO:0016020	3	0	48	18	8.09E-01	7.79E-02
Cellular Component: integral to membrane	GO:0016021	4	1	47	17	8.27E-01	5.96E-02
Molecular Function: transporter activity	GO:0005215	2	0	48	19	8.35E-01	5.12E-02
Biological Process: protein modification	GO:0006464	2	0	48	19	8.35E-01	5.12E-02
Biological Process: proteolysis	GO:0006508	0	5	43	21	8.47E-01	3.97E-02

attribute description	attribute	yy	ny	nn	yn	E_attribute	Gain
Biological Process: transport	GO:0006810	3	1	47	18	8.49E-01	3.76E-02
Biological Process: protein folding	GO:0006457	0	4	44	21	8.55E-01	3.15E-02
Molecular Function: oxidoreductase activity	GO:0016491	0	4	44	21	8.55E-01	3.15E-02
Molecular Function: calcium-dependent phospholipid binding	GO:0005544	1	0	48	20	8.61E-01	2.52E-02
Molecular Function: amidase activity	GO:0004040	1	0	48	20	8.61E-01	2.52E-02
Cellular Component: extracellular region	GO:0005576	1	0	48	20	8.61E-01	2.52E-02
Molecular Function: protein-L-isoaspartate [D-aspartate] O-methyltransferase activity	GO:0004719	1	0	48	20	8.61E-01	2.52E-02
Biological Process: exocytosis	GO:0006887	1	0	48	20	8.61E-01	2.52E-02
Biological Process: vesicle docking	GO:0048278	1	0	48	20	8.61E-01	2.52E-02
Molecular Function: structural constituent of ribosome	GO:0003735	1	0	48	20	8.61E-01	2.52E-02
Cellular Component: ribosome	GO:0005840	1	0	48	20	8.61E-01	2.52E-02
Molecular Function: molecular function unknown	GO:0005554	1	0	48	20	8.61E-01	2.52E-02
Biological Process: nucleotide metabolism	GO:0009117	1	0	48	20	8.61E-01	2.52E-02
Molecular Function: sugar porter activity	GO:0005351	1	0	48	20	8.61E-01	2.52E-02
Biological Process: carbohydrate transport	GO:0008643	1	0	48	20	8.61E-01	2.52E-02
Molecular Function: catalytic activity	GO:0003824	1	0	48	20	8.61E-01	2.52E-02
Molecular Function: nucleic acid binding	GO:0003676	1	0	48	20	8.61E-01	2.52E-02
Molecular Function: helicase activity	GO:0004386	1	0	48	20	8.61E-01	2.52E-02
Cellular Component: Golgi stack	GO:0005795	1	0	48	20	8.61E-01	2.52E-02
Biological Process: protein targeting	GO:0006605	1	0	48	20	8.61E-01	2.52E-02

Appendix I: Protein Domain Entropy Calculations

(only showing attributes with highest information gain)

Protein Domains- Entropy: Vesicles vs. Controls

79 vesicle proteins vs. 148 control proteins

Total Information Gain = 9.32E-01

attribute description	attribute	yy	ny	nn	yn	E_attribute	Gain
Chaperonin Cpn60/TCP-1	IPR002423	6	0	148	73	8.91E-01	4.12E-02
GroEL-like chaperone, ATPase	IPR008950	6	0	148	73	8.91E-01	4.12E-02
Chaperonin Cpn60	IPR001844	5	0	148	74	8.98E-01	3.42E-02
Chaperonin TCP-1	IPR002194	4	0	148	75	9.05E-01	2.73E-02
Histone core	IPR007125	3	0	148	76	9.12E-01	2.04E-02

Protein Domains- Entropy: Secretions vs. Controls

52 secretion proteins vs. 148 control proteins

Total Information Gain = 8.27E-01

attribute description	attribute	yy	ny	nn	yn	E_attribute	Gain
Histone core	IPR007125	3	0	148	49	7.97E-01	2.96E-02
Lactate/malate dehydrogenase	IPR001236	3	0	148	49	7.97E-01	2.96E-02
L-lactate/malate dehydrogenase	IPR001557	3	0	148	49	7.97E-01	2.96E-02
Protein kinase	IPR000719	0	9	139	52	8.07E-01	2.01E-02
Glutathione S-transferase, C-terminal	IPR004046	2	0	148	50	8.07E-01	1.96E-02
Malate dehydrogenase, active site	IPR001252	2	0	148	50	8.07E-01	1.96E-02

attribute description	attribute	yy	ny	nn	yn	E_attribute	Gain
Protein kinase-like	IPR011009	0	8	140	52	8.09E-01	1.78E-02
Histone-fold	IPR009072	3	1	147	49	8.11E-01	1.55E-02

Protein Domains- Entropy: Tegument vs. Controls

29 tegument proteins vs. 148 control proteins

Total Information Gain = 6.43E-01

attribute description	attribute	yy	ny	nn	yn	E_attribute	Gain
C2 calcium/lipid-binding region, CaLB	IPR008973	2	1	147	27	6.28E-01	1.58E-02
FerI	IPR012968	1	0	148	28	6.29E-01	1.49E-02
Allergen V5/Tpx-1 related	IPR001283	1	0	148	28	6.29E-01	1.49E-02
Aquaporin	IPR012269	1	0	148	28	6.29E-01	1.49E-02
Major intrinsic protein	IPR000425	1	0	148	28	6.29E-01	1.49E-02
Protein-L-isoaspartateD-aspartate O-methyltransferase	IPR000682	1	0	148	28	6.29E-01	1.49E-02
Exocyst complex component Sec10	IPR009976	1	0	148	28	6.29E-01	1.49E-02
Ribosomal protein L40e	IPR001975	1	0	148	28	6.29E-01	1.49E-02
Phosphatidylinositol transfer protein	IPR001666	1	0	148	28	6.29E-01	1.49E-02
Snf7	IPR005024	1	0	148	28	6.29E-01	1.49E-02
Spectrin repeat	IPR002017	1	0	148	28	6.29E-01	1.49E-02
Type I phosphodiesterase/nucleotide pyrophosphatase	IPR002591	1	0	148	28	6.29E-01	1.49E-02
PDZ/DHR/GLGF	IPR001478	1	0	148	28	6.29E-01	1.49E-02
General substrate transporter	IPR005828	1	0	148	28	6.29E-01	1.49E-02
Sugar transporter	IPR003663	1	0	148	28	6.29E-01	1.49E-02
AMP-dependent synthetase and ligase	IPR000873	1	0	148	28	6.29E-01	1.49E-02
DEAD/DEAH box helicase, N-terminal	IPR011545	1	0	148	28	6.29E-01	1.49E-02
Helicase, C-terminal	IPR001650	1	0	148	28	6.29E-01	1.49E-02
HSP20-like chaperone	IPR008978	1	0	148	28	6.29E-01	1.49E-02

attribute description	attribute	yy	ny	nn	yn	E_attribute	Gain
Heat shock protein Hsp20	IPR002068	1	0	148	28	6.29E-01	1.49E-02
Alpha crystallin	IPR001436	1	0	148	28	6.29E-01	1.49E-02
Protein of unknown function DUF1606	IPR011710	1	0	148	28	6.29E-01	1.49E-02
AP2 clathrin adaptor, alpha and beta chain, appendage	IPR009028	1	0	148	28	6.29E-01	1.49E-02
CAP protein	IPR001837	1	0	148	28	6.29E-01	1.49E-02

Protein Domains- Entropy: Secretions vs. Vesicles

53 secretion proteins vs. 78 vesicle proteins

Total Information Gain = 9.74E-01

attribute description	attribute	yy	ny	nn	yn	E_attribute	Gain
Chaperonin TCP-1	IPR002194	0	4	74	53	9.50E-01	2.33E-02
Chaperonin Cpn60/TCP-1	IPR002423	1	6	72	52	9.60E-01	1.32E-02
GroEL-like chaperone, ATPase	IPR008950	1	6	72	52	9.60E-01	1.32E-02
Tropomyosin	IPR000533	0	2	76	53	9.62E-01	1.15E-02
Myosin tail	IPR002928	0	2	76	53	9.62E-01	1.15E-02
Endoplasmic reticulum targeting sequence	IPR000886	0	2	76	53	9.62E-01	1.15E-02
Haloacid dehalogenase-like hydrolase	IPR005834	0	2	76	53	9.62E-01	1.15E-02
Cation transporting ATPase, N-terminal	IPR004014	0	2	76	53	9.62E-01	1.15E-02
ATPase, E1-E2 type	IPR001757	0	2	76	53	9.62E-01	1.15E-02
Cation transporting ATPase, C-terminal	IPR006068	0	2	76	53	9.62E-01	1.15E-02
E1-E2 ATPase-associated region	IPR008250	0	2	76	53	9.62E-01	1.15E-02

Protein Domains- Entropy: Tegument vs. Vesicles

29 tegument proteins vs. 78 vesicle proteins

Total Information Gain = 8.43E-01

attribute description	attribute	yy	ny	nn	yn	E_attribute	Gain
C2 calcium/lipid-binding region, CaLB	IPR008973	2	0	78	27	8.07E-01	3.59E-02
CD9/CD37/CD63 antigen	IPR000301	2	0	78	27	8.07E-01	3.59E-02
Tetraspanin	IPR008952	2	0	78	27	8.07E-01	3.59E-02
Chaperonin Cpn60/TCP-1	IPR002423	0	6	72	29	8.16E-01	2.65E-02
GroEL-like chaperone, ATPase	IPR008950	0	6	72	29	8.16E-01	2.65E-02
Chaperonin Cpn60	IPR001844	0	5	73	29	8.21E-01	2.19E-02
Annexin	IPR001464	1	0	78	28	8.25E-01	1.78E-02
Amidase	IPR000120	1	0	78	28	8.25E-01	1.78E-02
FerI	IPR012968	1	0	78	28	8.25E-01	1.78E-02
Allergen V5/Tpx-1 related	IPR001283	1	0	78	28	8.25E-01	1.78E-02
Aquaporin	IPR012269	1	0	78	28	8.25E-01	1.78E-02

Protein Domains- Entropy: Tegument vs. Secretions

30 tegument proteins vs. 51 secretion proteins

Total Information Gain = 9.51E-01

attribute description	attribute	yy	ny	nn	yn	E_attribute	Gain
C2 calcium/lipid-binding region, CaLB	IPR008973	2	0	51	28	9.15E-01	3.62E-02
Thioredoxin-like fold	IPR012336	2	0	51	28	9.15E-01	3.62E-02
CD9/CD37/CD63 antigen	IPR000301	2	0	51	28	9.15E-01	3.62E-02
Tetraspanin	IPR008952	2	0	51	28	9.15E-01	3.62E-02
Peptidase S1 and S6, chymotrypsin/Hap	IPR001254	0	3	48	30	9.26E-01	2.53E-02
Peptidase, trypsin-like serine and cysteine	IPR009003	0	3	48	30	9.26E-01	2.53E-02

attribute description	attribute	yy	ny	nn	yn	E_attribute	Gain
Peptidase S1A, chymotrypsin	IPR001314	0	3	48	30	9.26E-01	2.53E-02
Histone-fold	IPR009072	0	3	48	30	9.26E-01	2.53E-02
Histone core	IPR007125	0	3	48	30	9.26E-01	2.53E-02
Lactate/malate dehydrogenase	IPR001236	0	3	48	30	9.26E-01	2.53E-02
L-lactate/malate dehydrogenase	IPR001557	0	3	48	30	9.26E-01	2.53E-02
Phosphoglycerate mutase 1	IPR005952	1	0	51	29	9.33E-01	1.79E-02
Phosphoglycerate mutase	IPR013078	1	0	51	29	9.33E-01	1.79E-02

Appendix J: Perl And Matlab Scripts

```
#parse_nonred_seqs.pl
#This script was used in creating a control data set; given a file #containing a list of sequence ids
#corresponding entries that need to be eliminated from a fasta file, it will run through the fasta file
#and create two new files- one with eliminated sequences and one with
#the desired sequences

#usage: parse_nonredundant_seqs.pl

use strict;
use Bio::SeqIO;
open(INFILE, "s:\\sm_to_eliminate_new.txt");
local($/) = undef;
my $line = <INFILE>;

my $in  = Bio::SeqIO->new('-file' => "s:\\NRSm_455.fasta" ,
                                     '-format' => 'Fasta');
my $out = Bio::SeqIO->new('-file' => ">s:\\NRSm_455_eliminated.fasta" ,
                                      '-format' => 'Fasta');
my $out2 = Bio::SeqIO->new('-file' => ">s:\\NRSm_455_nonred.fasta" ,
                                      '-format' => 'Fasta');

while ( my $seq = $in->next_seq() ) {
    my $a = $seq->id;
    if ($line =~ /$a/){
        $out->write_seq($seq);}
     else {
      $out2->write_seq($seq); }}

#blastparser.pl
#written by Amy Schmidbauer
#This script will take a blast results file (standard long format) and parse
#specified results into table format
```

```
use strict;
use Bio::SearchIO;

my $outfile = "NRSm_155_blast_parsed.out";

open (OUTFILE, ">$outfile") or die "Can't open outfile";

    my $in = new Bio::SearchIO(-format => 'blast',
                               -file   => "NRSm_155_blast.out");
    while( my $result = $in->next_result ) {
      while( my $hit = $result->next_hit ) {
        while( my $hsp = $hit->next_hsp ) {
          if( $hsp->length('total') > 100 ) {
            #if ( $hsp->percent_identity >= 75 ) {
            if ($result->query_accession ne $hit->accession){
              print OUTFILE $result->query_accession,
                   ",", $result->query_description,
                   ",", $result->query_length,
                   ",", $result->num_hits,
                   ",", $hit->accession,
                   ",", $hit->description,
                     ",", $hit->significance,
                       ",", $hsp->length('total'),
                       ",", $hsp->percent_identity, "\n";
            }
          }
        }
      }
    }
```

```
#parse_iprscan.pl
#This script takes as input an InterPro scan file ("raw" format) and creates
#1- a matrix of seq_id vs. protein domain ids and descriptions
#2- a matrix of seq_id vs. GO ids and terms
#3- a file containing a count of seq ids corresponding to each protein domain
#4- a file containing a count of seq ids corresponding to each GO id
#could be split into 2 scripts or modularized for creating GO data vs. domain data

#input file must be sorted by protein_id

use Tie::IxHash;

open(INFILE1, "s:\\secretions_&_tegument_iprscan_cleaned.raw");
open(OUTFILE1, ">s:\\domains_matrix.out");
open(OUTFILE2, ">s:\\domain_id_counts.out");

my %HoH = ();
tie %HoH, "Tie::IxHash";
my %domain_ids = ();
tie %domain_ids, "Tie::IxHash";
my %domain_descr = ();
tie %domain_descr, "Tie::IxHash";

my $count=0;
#load protein->domain hash, domain_id->count hash,
#and domain_id->descr hash
while(<INFILE1>){
      if (/NULL/){
            @line = split("\t", $_);
            $prot_id = $line[0];
            $domain_id = "NULL";
            $HoH{$prot_id}->{$domain_id} = 1;}
      else  {
            @line = split("\t", $_);
            $prot_id      = $line[0];
            $domain_id    = $line[11];
            $domain_descr = $line[12];
            $domain_descr =~ tr/("|')//d;
            $HoH{$prot_id}->{$domain_id} = 1;
            $domain_ids{$domain_id} = $count;
            $domain_descr{$domain_id} = $domain_descr;}}

#print domain_ids as headers in domain_matrix output
printf OUTFILE1 '%20s'," ";
print OUTFILE1 "\t";
foreach my $a (keys %domain_ids){
      print OUTFILE1 "$a\t";  }
print OUTFILE1 "\n";

#print domain descriptions as headers in domain_matrix output
foreach my $a1 (keys %domain_descr)
    {
    my $domain_descr = $domain_descr{$a1};
    print OUTFILE1 "$domain_descr\t";
    }
print OUTFILE1 "\n";

#print protein->domain matrix data and load domain_id->count hash
foreach my $b (keys %HoH){
    #print $b;
    print OUTFILE1 "$b\t\t\t";
    foreach my $c (keys %domain_ids){
      #print "$c\n";
        if ($HoH{$b}->{$c}){
           #print "$b\t$c\n";
           print OUTFILE1 "1         \t";
           $domain_ids{$c} = ($domain_ids{$c} + 1);}
        else{
            #print "no $c\n";
            print OUTFILE1 "0         \t"; }}
    print OUTFILE1 "\n"; }
```

```
#print domain_id->description->count to output file
while ( my ($domain_id, $count) = each(%domain_ids) ){
    my $domain_descr = $domain_descr{$domain_id};
    print OUTFILE2 "$domain_id => $domain_descr => $count\n";}

close INFILE1;
close OUTFILE1;
close OUTFILE2;

open(INFILE1, "s:\\secretions_&_tegument_iprscan_cleaned.raw");
open(OUTFILE3, ">s:\\goids_matrix.out");
open(OUTFILE4, ">s:\\go_id_counts.out");

my %go_ids = ();  #hash
tie %go_ids, "Tie::IxHash";
my %go_descr = (); #hash
tie %go_descr, "Tie::IxHash";
my %prot_go = (); #hash of hashes
tie %prot_go, "Tie::IxHash";

my @line;
my $go_id;
my $size=0;
my $size2=0;
my @go_fields;
my $count2;
my $descr_long;
my $descr_go;
my $descr;

while(<INFILE1>) {
    @line = split("\t", $_);
    $size = @line;
    if ($size ==14){ #if GO column isn't empty
            $prot_id = $line[0];
            $go_line = $line[13];
            @go_fields = split(/\(/,$go_line);
            $size2 = @go_fields;
            for ($i=0; $i<$size2; $i++){
            if ($go_fields[$i] =~ /GO:\d{7}/){
                  #print "yes\n";
                  $go_id = $&;
                $count2 = 0;
                $descr_long = $go_fields[$i-1];
                if ($descr_long =~ /^GO:\d{7}/){
                  ($descr_go, $descr) = split(/,\s/,$descr_long);}
                else{
                  $descr = $descr_long;
                  $descr =~ tr/"//d;}
                #print "$descr\n\n";
                $go_ids{$go_id} = $count2;
                $go_descr{$go_id} = $descr;
                $prot_go{$prot_id} -> {$go_id} = 1;}}}
    else  #if GO column is empty{
            $prot_id = $line[0];
            $go_id = "NULL";
            $prot_go{$prot_id} -> {$go_id} = 1; }}

printf OUTFILE3 '%20s'," ";
print OUTFILE3 "\t";
foreach my $d (keys %go_ids){
      print OUTFILE3 "$d\t";}
print OUTFILE3 "\n";
printf OUTFILE3 '%20s'," ";
print OUTFILE3 "\t";
foreach my $e (keys %go_ids){
    my $go_descr = $go_descr{$e};
    print OUTFILE3 "$go_descr\t";}
print OUTFILE3 "\n";

#print protein->domain matrix data and load domain_id->count hash
foreach my $f (keys %prot_go){
    #print $f;
    print OUTFILE3 "$f\t\t\t";
    foreach my $g (keys %go_ids){
      #print "$g\n";
        if ($prot_go{$f}->{$g}){
            #print "yes $g\t$g\n";
            print OUTFILE3 "1         \t";
            $go_ids{$g} = ($go_ids{$g} + 1);}
        else{
            #print "no $f\t$g\n";
            print OUTFILE3 "0         \t"; } }
    print OUTFILE3 "\n"; }

#print domain_id->count hash
while ( ($go_id, $count2) = each(%go_ids )){
    my $go_descr = $go_descr{$go_id};
    print OUTFILE4 "$go_id => $go_descr => $count2\n"; }

close INFILE1;
close OUTFILE3;
close OUTFILE4;
```

```
#append_labels_to_matrix.pl
#Can be used to in conjunction with parse_iprscan.pl to prepare matrics for Matlab.
#Matrix files were openened in Excel so that a column containing class labels could
#be appended to the beginning of the matrix and all seq_ids and column headings could
#be deleted (so that matrix contained only 0s and 1s for use in Matlab).  Sometimes the
#matrix cannot be opened in Excel because it has too many columns.  This script is used to
#append a column of class labels to the beginning of the matrix.

use strict;

open(INFILE1, "s:\\secretions_&_tegument_domains_matrix.out");
open(OUTFILE1, ">s:\\secretions_&_tegument_domains_matrix2.out");

my $i = 1;
while(<INFILE1>){
      if ($i < 52){
            print OUTFILE1 "0\t$_";}
      else{
            print OUTFILE1 "1\t$_";}
      $i++; }

close INFILE1;
close OUTFILE1;
```

```
#Entropy.m
```

```
#Matlab script used to calculate Entropy

clear all

D = load('secretions_&_tegument_domains_matrix2.out');
fid = fopen('secretions_&_tegument_domain_entropy.txt','w');

fprintf(fid, '%s, %s, %s, %s\n', 'label_yes', 'label_no','total', 'I');
    label_yes = 0;
    label_no = 0;
    total = 0;
    for x = 1 : size(D,1)
        if (D(x,1)==1)
            label_yes = label_yes + 1;
        end
        if (D(x,1)==0)
            label_no = label_no + 1;
        end
    end

    if (label_yes + label_no) == 0
        I = 0;
    elseif (label_yes == 0 && label_no ~= 0)
        I = -(((label_no/(label_yes + label_no)) *log2(label_no)/(label_yes + label_no)));
    elseif (label_no == 0 && label_yes ~= 0)
        I = -(((label_yes/(label_yes + label_no))*log2(label_yes)/(label_yes + label_no)));
    else
        I = -((label_yes/(label_yes + label_no))*log2(label_yes/(label_yes + label_no)))-
((label_no/(label_yes + label_no))*log2(label_no/(label_yes + label_no)));
    end
    total = label_yes + label_no;
    fprintf(fid,'%d, %d, %d, %d\n', label_yes, label_no, total, I);

 fprintf(fid, '%s, %s, %s, %s, %s, %s, %s\n','attribute','yy', 'ny', 'nn', 'yn','E_attribute','Gain');
 for i = 2 : size(D,2)
    attribute = i-1;
    yy = 0;
    yn = 0;
    ny = 0;
    nn = 0;
    E_attribute = 0;
    Gain = 0;
    for x = 1 : size(D,1)
        if (D(x,1)==1) && (D(x,i)==1)
            yy = yy + 1;
        end
        if (D(x,1)==1) && (D(x,i)~=1)
            yn = yn + 1;
        end
        if (D(x,1)==0) && (D(x,i)==0)
            nn = nn + 1;
        end
        if (D(x,1)==0) && (D(x,i)~=0)
            ny = ny + 1;
        end
    end
    if total == 0
        E_attribute = 0;
    elseif ((yy == 0) && (ny ~= 0) && (nn ~= 0) && (yn ~= 0))
        E_attribute = (((ny+yy)/total) * (-(ny/(yy+ny))*log2(ny/(yy+ny)))) + (((yn+nn)/total) * (-
(yn/(yn+nn))*log2(yn/(yn+nn)) - (nn/(yn+nn))*log2(nn/(yn+nn))));
    elseif ((yy == 0) && (ny == 0) && (nn ~= 0) && (yn ~= 0))
        E_attribute = (((yn+nn)/total) * (-(yn/(yn+nn))*log2(yn/(yn+nn)) - (nn/(yn+nn))*log2(nn/(yn+nn))));
    elseif ((ny == 0) && (yy ~= 0) && (nn ~= 0) && (yn ~= 0))
        E_attribute = (((ny+yy)/total) * (-(yy/(yy+ny))*log2(yy/(yy+ny)))) + (((yn+nn)/total) * (-
(yn/(yn+nn))*log2(yn/(yn+nn)) - (nn/(yn+nn))*log2(nn/(yn+nn))));
    else
        E_attribute = (((ny+yy)/total) * (-(yy/(yy+ny))*log2(yy/(yy+ny)) - (ny/(yy+ny))*log2(ny/(yy+ny))))
+ (((yn+nn)/total) * (-(yn/(yn+nn))*log2(yn/(yn+nn)) - (nn/(yn+nn))*log2(nn/(yn+nn))));
    end
    Gain = I-E_attribute;
    fprintf(fid,'%d, %d, %d, %d, %d, %d, %d\n', attribute, yy, ny, nn, yn, E_attribute, Gain);
```

```
end

   fclose(fid);
```

Appendix K: Permission To Use Figures

Permission to use figures 2, 3, 4 (Dorsey, Cousin et al. 2002):

From: "Jones, Jennifer (ELS-OXF)" <J.Jones@elsevier.co.uk>
To: "'schmidbauer2@comcast.net'" <schmidbauer2@comcast.net>
Subject: RE: Obtain Permission
Date: Monday, June 05, 2006 3:50:26 AM

Dear Amy Schmidbauer

We hereby grant you permission to reproduce the material detailed below at no charge in your thesis subject to the following conditions:

1. If any part of the material to be used (for example, figures) has appeared in our publication with credit or acknowledgement to another source, permission must also be sought from that source. If such permission is not obtained then that material may not be included in your publication/copies.

2. Suitable acknowledgment to the source must be made, either as a footnote or in a reference list at the end of your publication, as follows:

"Reprinted from Publication title, Vol number, Author(s), Title of article, Pages No., Copyright (Year), with permission from Elsevier".

3. Reproduction of this material is confined to the purpose for which permission is hereby given.

4. This permission is granted for non-exclusive world English rights only. For other languages please reapply separately for each one required. Permission excludes use in an electronic form. Should you have a specific electronic project in mind please reapply for permission.

5. This includes permission for UMI to supply single copies, on demand, of the complete thesis. Should your thesis be published commercially, please reapply for permission.

Yours sincerely

Jennifer Jones

Rights Assistant

-----Original Message-----

From: schmidbauer2@comcast.net [mailto:schmidbauer2@comcast.net]
Sent: 04 June 2006 23:27
To: permissions@elsevier.com
Subject: Obtain Permission

This Email was sent from the Elsevier Corporate Web Site and is related to Obtain Permission form:

Product: Customer Support
Component: Obtain Permission
Web server: http://www.elsevier.com
IP address: 10.10.24.149
Client: Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1; .NET CLR 1.0.3705; .NET CLR 1.1.4322)
Invoked from:
http://www.elsevier.com/wps/find/obtainpermissionform.cws_home?isSubmitted=yes&navigateXmlFileName=/store/prod_webcache_act/framework_support/obtainpermission.xml
Request From:
Ms Amy Schmidbauer
Indiana University Purdue University Indianapolis
4720 Stoughton Ct.
46254
Indianapolis
United States
Contact Details:
Telephone: 317-250-3071
Fax:
Email Address: schmidbauer2@comcast.net
To use the following material:
ISSN/ISBN:
Title: Micron
Author(s): Dorsey, C., Cousin, C., Lewis, F., Stirewalt, M.
Volume: 33
Issue: 3
Year: 2002
Pages: 279 - 323
Article title: Ultrastructure of the *S. mansoni* cercaria
How much of the requested material is to be used: Figures 1, 10, and 23a
Are you the author: No
Author at institute: No
How/where will the requested material be used:

In a graduate thesis through IUPUI on cercarial secretions of the parasite *S. mansoni*.

For further info regarding this automatic email, please contact:
WEB APPLICATIONS TEAM (esweb.admin@elsevier.co.uk)

From: "Cousin, Carolyn" <ccousin@udc.edu>
To: <schmidbauer2@comcast.net>
Subject: RE: RE: permission to use figure
Date: Tuesday, May 16, 2006 12:22:23 PM

I would be willing to grant permission for you to use the requested figures from the publication. If you need my signature on a letter, I can supply it via e-mail.

Best of luck on your project and manuscript.

Carolyn Cousin

-----Original Message-----

From: schmidbauer2@comcast.net [mailto:schmidbauer2@comcast.net]
Sent: Tuesday, May 16, 2006 10:28 AM
To: Cousin, Carolyn
Subject: FW: RE: permission to use figure

Dear Dr. Cousin,

Dr. Fred Lewis advised me to seek your permission, in addition to his and the publisher's, to use figures 1, 10, and 23a from the publication "Ultrastructure of the *S. mansoni* cercaria" (Micron Vol. 33, pages 279-323). I am doing a thesis project at Indiana University that is a proteomic study of the *S. mansoni* secretome. Would you be willing to grant this permission?
thank you,

Amy Schmidbauer

-----Original Message-----

From: "Fred Lewis" <flewis@afbr-bri.com>
To: <schmidbauer2@comcast.net>
Subject: RE: permission to use figure
Date: Tuesday, May 16, 2006 9:21:39 AM

Amy,

You have my permission to use those figures, but I may not have the final say on the subject. You will also need to contact the publisher, Pergamon Press, which produces the journal Micron in which the article appears. Although unlikely, there may be some copyright issues involved. Also, please contact Carolyn Cousin at ccousin@udc.edu for her consent as well. Dr. Stirewalt is no longer alive, and Charles Dorsey is in very ill health.

Fred

Fred Lewis, Ph.D.
Head, Schistosomiasis Laboratory
Biomedical Research Institute
12111 Parklawn Dr, Rockville MD 20852
phone (301)881-3300 ext 25 or ext 27
fax (301) 770-4756

-----Original Message-----

From: schmidbauer2@comcast.net [mailto:schmidbauer2@comcast.net]
Sent: Monday, May 15, 2006 5:12 PM
To: flewis@afbr-bri.com
Subject: permission to use figure

Dear Dr. Lewis,

I am doing a thesis project at Indiana University that is a proteomic study of the *S. mansoni* secretome. I am writing to seek permission to use figures 1, 10, and 23a from the publication "Ultrastructure of the *S. mansoni* cercaria" (Micron Vol. 33, pages 279-323) in my thesis. This paper was extremely helpful to me in gaining an understanding of the morphology of this parasite. Would you be willing to grant this permission?

thank you,

Amy Schmidbauer

Permission to use figure 5 (Jones, Gobert et al. 2004):

From: "Malcolm Jones" Malcolm.Jones@qimr.edu.au
To: schmidbauer2@comcast.net

Subject: Re: permission to use figure
Date: Sunday, May 14, 2006 5:42:45 PM

Dear Ms Schmidbauer,
Very happy to grant the permission. Thanks indeed for asking. Would you prefer the original digital image, or will you make a copy from the published article?

Best wishes

Malcolm

-----Original Message-----

From: schmidbauer2@comcast.net
To: malcolmJ@qimr.edu.au
Subject: permission to use figure
Date: Sunday, May 14, 2006 10:11:31 AM

Dear Dr. Malcolm Jones,

I am doing a thesis project at Indiana University that is a proteomic study of the *S. mansoni* secretome. I am writing to seek permission to use a figure from one of your publication in my thesis- Figure 2 from "The cytoskeleton and motor proteins of human schistosomes and their role in surface maintenance and host-parasite interactions" from Bioessays Vol. 26 No. 7 pages 752-765 2004.

thank you,

Amy Schmidbauer

REFERENCES

Aitkin, M. and D. B. Rubin (1985). "Estimation and hypothesis testing in finite mixture models." Journal of the Royal Statistical Society **B**(47): 67–75.

Alberts, B., D. Bray, et al. (1994). Molecular Biology of the Cell, 3rd Edition. New York City, Garland Publishing.

Altschul, S. F., T. L. Madden, et al. (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs " Nucleic Acids Research **25**: 3389-3402.

Bailey, T. L. and C. Elkan (1994). "Fitting a mixture model by expectation maximization to discover motifs in biopolymers." Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology: 28-36.

Bailey, T. L. and M. Gribskov (1998). "Combining evidence using p-values: Application to sequence homology searches." Bioinformatics **14**: 48-54.

Bairoch, A. and B. Boeckmann (1994). "The SWISS-PROT protein sequence data bank: current status." Nucleic Acids Research **22**: 3578-3580.

Bendtsen, J. D., H. Nielsen, et al. (2004). "Improved prediction of signal peptides: SignalP 3.0." Journal of Molecular Biology **340**: 783-795.

Bergquist, N. R. (1998). "Schistosomiasis vaccine development: progress and prospects. ." Mem. Inst. Oswaldo Cruz **93** (suppl.1): 95-101.

Bergquist, R. (2004). Prospects for schistosomiasis vaccine development. T. D. R. News. **71**.

Bhattacharjee, S., N. L. Hiller, et al. (2006). "The Malarial Host-Targeting Signal Is Conserved in the Irish Potato Famine Pathogen." PLOS Pathogens **2**(5): 453-465.

Brindley, P. J. and T. P. Yoshino (2003). "Mobile genetic elements colonizing the genomes of metazoan parasites." Trends in Parasitology **18**: 79-87.

Capron, M. and A. Capron (1994). "Immunoglobulin E and Effector Cells in Schistosomiasis." Science **264**(5167): 1876-1877.

Chitsulo, L. (2005). Schistosomiasis scientific working group. TDR News.

Chitsulo, L., D. Engels, et al. (2000). "The global status of schistosomiasis and its control." Acta Trop. **77**(1): 41-51.

Cioli, D. and L. Pica-Mattoccia (2005). Schistosomiasis (Chapter 13. Current and Future Anti-Schistosomal Drugs), Springer.

Cioli, D., L. Pica-Mattoccia, et al. (1992). "Schistosoma mansoni: Hycanthone/oxamniquine resistance is controlled by a single autosomal recessive gene." Experimental Parasitology **75**: 425-432.

Clamp, M., J. Cuff, et al. (2004). "The Jalview Java Alignment Editor." Bioinformatics **20**: 426-7

Crooks, G. E., G. Hon, et al. (2004). "WebLogo: A Sequence Logo Generator." Genome Research **14**: 1188-1190.

Curwen, R. S., P. D. Ashton, et al. (2004). "The Schistosoma mansoni soluble proteome: a comparison across four life-cycle stages." Mol. Biochem. Parasitol. **138**(1): 57-66.

Doenhoff, M. J., A. A. Sabah, et al. (1987). "Evidence for an immune-dependent action of praziquantel on *Schistosoma mansoni* in mice." Trans. R. Soc. Trop. Med. Hyg. **81**: 947-951.

Dorsey, C. H., C. E. Cousin, et al. (2002). "Ultrastructure of the *Schistosoma mansoni* cercaria." Micron **33**: 279-323.

Duckert, P., S. Brunak, et al. (2004). "Prediction of proprotein convertase cleavage sites." Protein Engineering, Design & Selection **17**(1): 107-112.

Edman, J. C., L. Ellis, et al. (1985). "Sequence of protein disulphide isomerase and implications of its relationship to thioredoxin." Nature **317**(6034): 267-70.

El-Sayed, N. M. A., D. Bartholomeu, et al. (2004). "Advances in Schistosome genetics." Trends in Parasitology **20**(4).

Fetterer, R. H., R. A. Pax, et al. (1980). "Praziquantel, potassium and 2,4-dinitrophenol: analysis of their action on the musculature of Schistosoma mansoni." European Journal of Pharmacology **64**: 31-38.

Fietto, J. L. R., R. DeMarco, et al. (2002). "Use of degenerate primers and touchdown PCR for construction of cDNA libraries." Biotechnology **32**: 1404-1411.

Fitzpatrick, J. M., D. A. Johnston, et al. (2005). "An oligonucleotide microarray for transcriptome analysis of Schistosoma mansoni and its application/use to investigate gender-associated gene expression." Molecular & Biochemical Parasitology **141**(1): 1-13.

Fogel, G. B., D. G. Weekes, et al. (2004). "Discovery of sequence motifs related to coexpression of genes using evolutionary computation." Nucleic Acids Research **32**(13): 3826–3835.

Freedman R. B., Hirst T. R., et al. (1994). "Protein disulphide isomerase: building bridges in protein folding." Trends Biochem Sci. **19**(8): 331-336.

Fusco, A. C., B. Salafsky, et al. (1991). "Schistosoma mansoni: the role of calcium in the stimulation of cercarial proteinase release." J Parasitol. **77**(5): 649-57

Gonnert, R. and P. Andrews (1977). "Praziquantel, a new broad-spectrum antischistosomal agent." Z. Parasitenk **52**: 129-150.

Haldar, K., N. L. Hiller, et al. (2005). "*Plasmodium* parasite proteins and the infected erythrocyte." Trends in Parasitology **21**(9).

Harder, A., J. Goossens, et al. (1988). "Influence of praziquantel and calcium on the bilayer-isotropic-hexagonal transition of model membranes." Mol. Biochem. Parasitol. **29**: 55-60.

Heijne, G. (1986). "A new method for predicting signal sequence cleavage sites." Nucl. Acids Res **14**: 4683-4690.

Hertz, G. and G. Stormo (1999). "Identifying DNA and protein patterns with statistically significant alignments of multiple sequences." Bioinformatics **15(7-8)**(Jul-Aug): 563-77.

Hiller, N. L., S. Bhattacharjee, et al. (2004). "A Host-Targeting Signal in Virulence Proteins Reveals a Secretome in Malarial Infection." Science **306**(5703): 1934-1937.

Hillyer, G. V. (1974). "Buoyant density and thermal denaturation profiles of schistosome DNA." Journal of Parasitology **60**: 725-727.

Hoffmann, K. F., D. A. Johnston, et al. (2002). "Identification of Schistosoma mansoni gender-associated gene transcripts by cDNA microarray profiling." Genome Biology **3**(8): research0041.1-0041.12.

Hu, W., Q. Yan, et al. (2003). "Evolutionary and biomedical implications of a Schistosoma japonicum complementary DNA resource." Nature Genetics **35**(2): 139-147.

Hua, S. and Z. Sun (2001). "Support vector machine approach for protein subcellular localization prediction." Bioinformatics **17**: 721-728.

Iseli, C., C. V. Jongeneel, et al. (1999). "ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences." Proc Int Conf Intell Syst Mol Biol. : 138-148.

Johnston, D. A., M. Blaxter, et al. (1999). "Genomics and the biology of parasites." BioEssays **21**: 131–147.

Jones, M. K., G. N. Gobert, et al. (2004). "The cytoskeleton and motor proteins of human schistosomes and their roles in surface maintenance and host-parasite interactions." Bioessays **26**(7): 752-765.

Juncker, A. S., H. Willenbrock, et al. (2003). "Prediction of lipoprotein signal peptides in Gram-negative bacteria." Protein Sci. **12**: 1652-1662.

Karanja, D. M., A. E. Boyer, et al. (1998). "Studies on schistosomiasis in western Kenya: II. Efficacy of praziquantel for treatment of schistosomiasis in persons coinfected with human immunodeficiency virus-1." Am. J. Trop. Med. Hyg. **59**: 307-311.

Katz, N., E. P. Dias, et al. (1973). "Estudo de uma cepa humana de *Schistosoma mansoni* resistente a agentes esquistossomicidas." Rev. Soc. Bras. Med. Trop. **7**: 381-387.

Knudsen, G. M., K. F. Medzihradszky, et al. (2005). "Proteomic Analysis of *Schistosoma mansoni* Cercarial Secretions." Molecular & Cellular Proteomics **4**: 1862-1875.

Kohn, A. B., P. A. Anderson, et al. (2001). "Schistosome calcium channel beta subunits. Unusual modulatory effects and potential role in the action of the antischistosomal drug praziquantel." J. Biol. Chem. **276**: 36873-36876.

Le Paslier, M. C., R. J. Pierce, et al. (2000). "Construction and characterization of a *Schistosoma mansoni* bacterial artificial chromosome library." Genomics **65**: 87-94.

Le, T. H., D. Blair, et al. (2002). "Mitochondrial genomes of parasitic flatworms." Trends in Parasitology **18**: 206-213.

Lingelbach, K. and K. A. Joiner (1998). "The parasitophorous vacuole membrane surrounding Plasmodium and Toxoplasma: an unusual compartment in infected cells." Journal of Cell Science **111**(11): 1467-1475.

Liu, X., D. Brutlag, et al. (2001). BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes Pacific Symposium on Biocomputing.

Loker, E. S. and G. M. Mkoji (2005). Schistosomiasis (Chapter 1. Schistosomes and Their Snail Hosts), Springer.

Lopez-Estraño, C., S. Bhattacharjee, et al. (2003). "Cooperative domains define a unique host cell-targeting signal in *Plasmodium falciparum*-infected erythrocytes." Proceedings of the National Academy of Science U.S.A **100**(21): 12402-102407.

Marti, M., R. T. Good, et al. (2004). "Targeting Malarial Virulence and Remodeling Proteins to the Host Erythrocyte." Science **306**(5703): 1930-1933.

McLaren, D. J. and D. J. Hockley (1977). "Blood flukes have a double outer membrane." Nature **269**: 147-149.

McTigue, M. A., D. R. Williams, et al. (1995). "Crystal structure of a schistosomal drug and vaccine target: glutathione S-transferase from *Schistosoma japonica* and its complex with the leading antischistosomal drug praziquantel." J. Mol. Biol. **246**: 21-27.

Mountford, A. and S. Jenkins (2005). Schistosomiasis (Chapter 5. Vaccine Development), Springer.

Mulder, N. J., R. Apweiler, et al. (2005). "InterPro, progress and status in 2005." Nucleic Acids Res. **33 (Database Issue)**: 201-5.

Nakai, K. and P. Horton (1999). "PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. ." Trends Biochem. Sci. **24**: 34-36.

Nechay, B. R., G. R. Hillman, et al. (1980). "Properties and drug sensitivity of adenosine triphosphatases from *Schistosoma mansoni*." Journal of Parasitology **66**: 596-600.

Nielsen, H., J. Engelbrecht, et al. (1997). "Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites." Protein Engineering **10**(1): 1-6.

Ohler, U. and H. Niemann (2001). "Identification and analysis of eukaryotic promoters: recent computational approaches." Trends in Genetics **17**(2): 56-60.

Pax, R., J. L. Bettett, et al. (1978). "A benzodiazepine derivative and praziquantel: Effects on musculature of *Schistosoma mansoni* and *Schistosoma japonicum*." Naunyn-Schiedberg's Arch. Pharmacol. **304**: 309-315.

Pearson, M. S., D. P. McManus, et al. (2005). "In vitro and in silico analysis of signal peptides from the human blood fluke, *Schistosoma mansoni*." FEMS Immunology and Medical Microbiology **45**: 201-211.

Pica-Mattoccia, L., L. C. S. Dias, et al. (1993). "Schistosoma mansoni: Genetic complementation analysis shows that two independent hycanthone/oxamniquine-resistant strains are mutated in the same gene." Experimental Parasitology **77**: 445-449.

Price, H., M. Doenhoff, et al. (1997). "Cloning, heterologous expression and antigenicity of a schistosome cercarial protease." Parasitology **114**: 447-453.

Quackenbush, J., F. Liang, et al. (2000). "The TIGR Gene Indices: reconstruction and representation of expressed gene sequences." Nucleic Acids Research **28**: 141-145.

Remme, J., E. Blas, et al. (2002). "Strategic emphasis for tropical disease research: a TDR perspective." Trends in Parasitology **18**(10): 421-426.

Rogers, S. H. and E. Bueding (1971). "Hycanthone resistance: Development in *Schistosoma mansoni*." Science **172**: 1057-1058.

Sabah, A. A., C. Fletcher, et al. (1985). "Schistosom mansoni: reduced efficacy of chemotherapy in infected T-cell-deprived mice." Experimental Parasitology **60**: 348-354.

Sauma, S. Y. and M. Strand (1990). "Identification and characterization of glycophosphatidylinositol-linked Schistosoma mansoni adult worm immunogens." Mol. Biochem. Parasitol. **38**: 199-210.

Schepers, H., R. Brasseur, et al. (1988). "Mode of insertion of praziquantel and derivatives into lipid membranes." Biochem. Pharmacol. **37**: 1615-1623.

Schmidt, G. D. and L. S. Roberts (2000). Foundations of Parasitology, 6th edition. Chapter 15. Trematoda: Form, Function, and Classification of Digeans McGraw-Hill Comp.

Short, R. B., J. D. Liberatos, et al. (1989). "Conventional Giesmsa-stained and C-banded chromosomes of seven strains of *Schistosoma mansoni*." Journal of Parasitology **75**: 920-926.

Simpson, A. J., A. Sher, et al. (1982). "The genome of Schistosoma mansoni: isolation of DNA, its size, bases and repetitive sequences." Mol. Biochem. Parasitol. **6**: 125-137.

Slater, G. (1996-1999). Expressed Sequence Tag Analysis Tools Etc (c) Human Genome Mapping Project RC. Hinxton, Cambridge, UK.

Smyth, D., D. P. McManus, et al. (2003). "Isolation of cDNAs Encoding Secreted and Transmembrane Proteins from *S. mansoni* by a Signal Sequence Trap Method." Infection and Immunity **71**(5): 2548-2554.

Stirewalt, M. A. and F. J. Kruidenier (1961). "Activity of the acetabular secretory apparatus of cercariae of Schistosoma mansoni under experimental conditions. ." Experimental Parasitology **11**: 191–211.

Takemoto, H., T. Yoshimori, et al. (1992). "Heavy chain binding protein (BiP/GRP78) and endoplasmin are exported from the endoplasmic reticulum in rat exocrine pancreatic cells, similar to protein disulfide-isomerase." Archives of biochemistry and biophysics **296**(1): 129-36.

TDR (2002). TDR Strategic Direction for Research: Schistosomiasis. World Health Organization Tropical Disease Research Unit Strategic Direction, World Health Organization Tropical Disease Research Unit.

Thompson, J. D., D. G. Higgins, et al. (1994). "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. ." Nucleic Acids Research **22**: 4673-4680.

van Balkom, B. W. M., R. A. van Gestel, et al. (2005). "Mass Spectrometric Analysis of the Schistosoma mansoni Tegumental Sub-proteome." Journal of Proteome Research **4**(3): 958-66.

Verjovski-Almeida, S., R. DeMarco, et al. (2003). "Transcriptome analysis of the acoelomate human parasite Schistosoma mansoni." Nat. Genet.. **35**(2): 148-157.

Wickham, M. E. (2001). "Trafficking and assembly of the cytoadherence complex in *Plasmodium falciparum*-infected human erythrocytes." EMBO J **20**(20): 5536-5649.

Williams, D. L. and R. J. Pierce (2005). Schistosomiasis (Chapter 4. *Schistosoma* Genomics), Springer.

Yuan, X., J. Shen, et al. (2005). "Schistosoma japonicum: a method for transformation by electroporation." Exp Parasitol. **111**(4): 244-9.

CURRICULUM VITAE

CONTACT INFORMATION

Amy L. Schmidbauer
4720 Stoughton Court
Indianapolis, IN 46254
317-293-7821
317-250-3071 (cell)
aschmidb@iupui.edu
alschmidbauer@dow.com

EDUCATION

2002-2006	Indiana University Purdue University Indianapolis M.S. Bioinformatics (expected)
1998	International Society of Seed Technologists R.S.T. (Registered Seed Technologist)
1989-1993	University of Illinois at Urbana-Champaign B.S. Microbiology (Major: Microbiology, Minor: Chemistry)

PROFESSIONAL AFFILIATIONS

International Society for Computational Biology (ISCB), 2005-present
Symposium on Applied Computing Machinery, Bioinformatics Track Reviewer 2005-7
Crop Science Society of America

PUBLICATIONS

Indiana Bioinformatics Conference, 2006, Poster: “Information Theoretic Approaches to Analyzing Protein Domain and Gene Ontology Annotation Data: A Study of *Schistosoma mansoni*”

PROFESSIONAL EXPERIENCE

June 2006- Present **Bioinformatics Specialist**
Dow AgroSciences, Indianapolis, IN
Discovery Research Information Management

1. Lead the development of bioinformatics strategy.
2. Manage bioinformatics strategy implementation projects.
3. Lead external collaborations and intern projects.
4. Develop expertise in Pipeline Pilot and develop automated bioinformatics workflows.

Oct 2001- June 2006 **Business & Systems Analyst**
Dow AgroSciences, Indianapolis, IN
Discovery Research Information Management

1. Act as liaison between research community and Dow IS.
2. Act as partner project manager for biotech-related IS projects including process mapping, business requirements analysis, solution analysis, application testing and implementation, and user training and support.
3. Participate on Six-Sigma and Design for Six-Sigma projects for process improvement or new process development related to recombinant and native culture collections.
4. Provide informatics support for Microbiology and Natural Products groups, including query/report development and data mining support.
5. Provide commercial software evaluations, comparisons, and recommendations based on researchers' needs.
6. Subject Matter Expert for Spotfire. Lead monthly Spotfire user group, provide user training and support, perform software upgrades, configure guides and tools.
7. Subject Matter Expert for Nautilus LIMS. Lead a subject matter expert group, provide workflow configuration, user support and training, application troubleshoooting.
8. Develop and maintain Discovery and Bioinformatics websites.

Apr 2001-Oct 2001 **LIMS Analyst**
Dow AgroSciences, Indianapolis, IN
Bioinformatics Group, Genomics

1. Develop expertise in Nautilus LIMS to enable high-throughput screening processes.
2. Map laboratory processes and develop corresponding workflows in Nautilus LIMS to track laboratory processes and data.
3. Provide Nautilus troubleshooting, user support and training.
4. Perform software upgrades as necessary.

Mar 1998- Mar 2001 **Assistant Lab Manager**
Cargill Hybrid Seeds
Aurora, IL
Quality Assurance

1. Supervise daily lab activities.
2. Provide process and lab improvements to increase efficiency and effectiveness and meet tight deadlines.
3. Provide employee development opportunities and performance evaluations.
4. Make decisions to certify or fail commercial seed lots for shipment.
5. Manage large-scale project to test and certify all non-GMO seed lots as non-GMO.
6. Mentor associate at Syngenta Seeds in studying for RST exam.

Mar 1996- Mar 1998 **Seed Technologist**
Cargill Hybrid Seeds
Aurora, IL
Quality Assurance

1. Develop expertise in International Seed Testing Association (ISTA) rules and procedures.
2. Ensure seed quality is assessed according to ISTA rules and seed quality matches official lot tags.
3. Develop new high-throughput analytical methods for testing genetic insect and herbicide resistance traits.
4. Perform high-throughput germination, purity, and trait testing.